To our son, Ethan

Preface

Many statistics books begin with the elementary notions of statistics such as mean, standard deviation, etc. and end with regression or classification models (Refs??). Others start with regression and/or classification and delve deeply into their intricacies (Refs??). This book will begin with the topics of regression and classification, in their traditional guise, but then take an orthogonal direction by exploring the same topics in the context of neural networks. The need for such a book has arised because of the recently gained popularity of neural networks among applied statisticians (refs??). However, most neural network books are written by and for computer scientists and engineers, and at a level of mathematical rigor that is somewhat beyond the interest of the "average" social scientist interested in neural networks as a tool for analyzing statistical data. This book attempts to fill that void by bridging the gap between regression and classification, as employed in the social sciences, and neural networks. But, what is a neural network?

There exists a plethora of definitions for a neural network; however, it is safe to say that at least one type of neural network is a statistical tool that resembles regression and classification techniques. It is often said that neural networks have certain advantages over other statistical tools that render them "superior" to other methods. The current belief, as well as that of the authors, maintains that though neural networks are not a panacea, the advantages are sufficiently significant to justify their inclusion into the exisiting paraphernalia of statistical tools.

To further suite the social sciences, almost the first half of the book deals with topics that are part of the upbringing of any social scientist, namely regression and classification. This serves two functions: first, to set down the notation and formalism that will be employed in the second half of the book (the neural network-related chapters), and second, to provide for an arena wherein the reader is surrounded by an air of familiarity.

Our intent to reach the social scientist has made it difficult to decide on the level of mathematical rigor and complexity of this book. This is so because of the wide range of mathematical expertise in the social sciences. This variation exists among students and researchers alike. The way in which we have remedied that problem

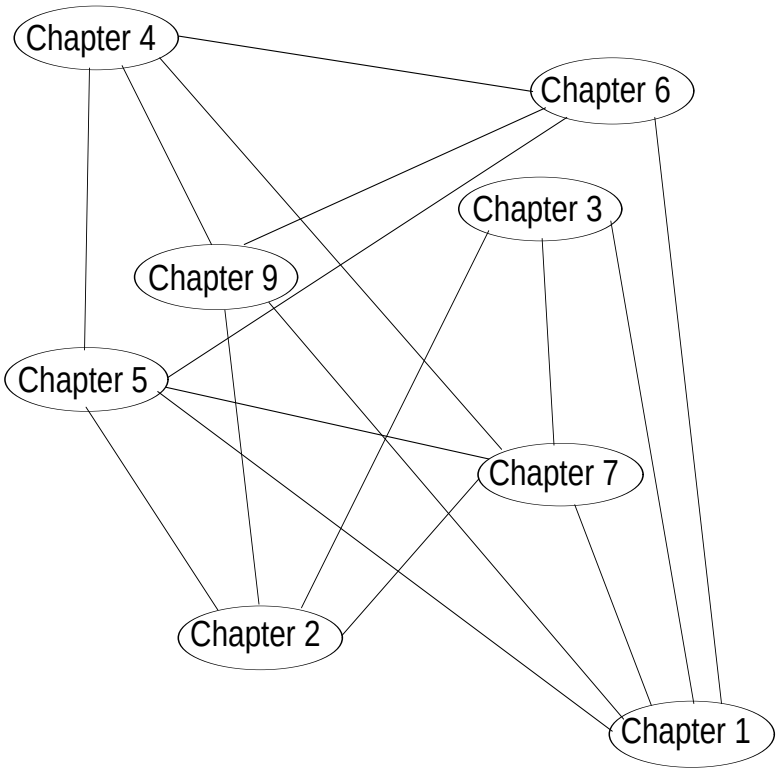


Fig. 0.1 The ideal structure of an imaginary book.

is by first presenting the relevant mathematical equations which may be neglected by the non-mathematical reader - student or researcher - and second, by describing the equations in pictorial terms, pictures, graphs, plots, etc..

Finally, the matter of the structure of the book: All books are sequential in that there is a natural starting and ending point (with the exception of dictionaries). However, rarely is a body of knowledge sequential. Even the historical development of most topics is highly convoluted, with a great deal of back-tracks and futile attempts. Anyone who has written a book, a thesis, a dissertation, or any large-scale compilation will attest to the difficulty of reducing such an inherently non-sequential structure to an explicitly sequential one. Furthermore, any reduction is apt to yield a work in which the various chapters are not mutually exclusive; the first chapter may be naturally overlapping with the last one, and a middle chapter may be intimately related to, and even based on, several of the later chapters. Indeed, a faithful structure for any book would resemble Figure ??. This book is no exception, and in fact we shall take this structure as our first encounter with a neural network structure.

Contents

Chapter 1
Background

1.1 Introduction

This book is intended for the social scientist. However, given the labyrinthine history of neural networks (NN), it behooves one to at least be aware of the origins and the history of the field. A brief history is as follows.

It may not come as a surprise that the origins of NNs lay in neuroscience (Kohonen ...??). The idea was to develop a model of the brain that can not only simulate everyday behavior, such as the escape of a rabbit at the sight (??) of a lion, but also explain some of the less mundane aspects of cognition, such as consciousness and the idea of inner-self (Chalmers, D. J. 1996; The Conscious Mind: IN Search of a Fundamental Theory, NY: Oxford University Press.??)

1.1.1 *The Brain*

Prior to 1880 the nervous system was believed to be a continuous ....

Then, Camillo Golgi invented a method for staining nerve fibers ....

In 1888, Santiago Ramon y Cajal employed this technique and showed that the nervous system was in fact "quantized" in that there exist tiny gaps ($\sim 200nm$) between individual neurons ....

In 1906, Golgi and Cajal shared the Nobel prize for their discoveries ....

Review the NN models of the brain (Amit, 1989) and the computational brain (Churchland and Sejnowski, 1992). Include some recent articles.

1.1.2 *Physics*

Understanding the behavior of a system of interacting particles is a standard problem in many branches of physics. It was, therefore, only natural for physicists to attack the problem of modeling the brain with their own bag of tricks.

Of course, no physicist has dared to claim that the problem has been solved; however, several "toy models" have been successfully solved, shedding some light

on the inner-workings of the brain.

An extremely simple model of the brain - at least of the portions dealing with memory - is offered by a system of bar-magnets! ...

A more sophisticated model is called a *spin glass* and ....

Review Muller and Reinhardt (1991).

1.1.3 *Society*

Webster's Ninth Collegiate Dictionary defines "society" as: "An enduring and cooperating social group whose members have developed organized patterns of relationships through interactions with one another; an interdependent system of organisms or biological units."

The similarities of the brain and society are too striking to have been overlooked in the past (refs??). However, in spite the inherent similarities, the methodological gap between brain science and the social sciences would be too wide to close were it not for the advent of Neural Networks as a scientific discipline. This field has brought together a diverse mix of otherwise unrelated disciplines .... This act has been called a "revolution" by some (refs ??).

One system of extreme complexity that has not yet fully benefited from this revolution is human society itself. Indeed, as is usually the case with novel approaches, the introduction of NNs into the social sciences has met with a certain degree of resistance....

1.2 NNs and Statistics

As seen from the previous discussion, an NN generally refers to a network of elementary processing units, called neurons or *nodes*, interconnected via synaptic connections, or simply, *weights*. It is the set of values assigned to these weights and the way in which they interconnect that determines the task the NN is to perform.

Most NNs do one (or both) of two things, regression and/or classification. When one can talk about an *input node*, it is equivalent to an independent variable and can be thought of as a predictor, or a regressor. If there exists a node that can be called an *output node*, then it can be thought of as a dependent or a response variable. Input nodes do not receive a signal from other nodes; their input comes from some external stimulus. Output nodes, by contrast, have no output signal into other nodes; they are at the end of a chain of computations performed by a number of "prior" nodes. Such an NN, or more accurately, the way in which it is trained is called *Supervised*.

To arrive at the desired value for the weights one must *train* a NN. ... The output of a NN may not necessarily be exactly what one desires. The desired value is called the *target*, and training is the process of assuring that the outputs are as close as possible to the corresponding targets. Therefore, a statistician would

recognize a NN as a parametric model, and training would be nothing but parameter estimation.

In another sense, NNs are not parametric models. ... Later.

Talk about (in words) about NNs for cluster analysis.

Although, it is true that a certain class of NNs are capable of approximating "any" function to any desired accuracy (Hornik, Stinchcombe, and White, 1989), the same is also true of many traditional methods, such as spline regression, polynomial regression, and projection pursuit (Sarle 1994a). Therefore, NNs are not to be considered a panacea. If there is any advantage that NNs have over other methods it is in the way in which they handle the problem of "the curse of dimensionality." Briefly, the number of free parameters in polynomial regression, for example, increases exponentially with the number of independent variables. By contrast, the number of free parameters in an NN (with one hidden layer) grows only linearly. The "explosion" of the number of free parameters in polynomial regression makes it more prone to overfitting problems. On the other hand, the drastically smaller number of parameters in the NN renders it less likely to overfit data, yet it does not prevent it from approximating "any" function. All of this will be discussed in more detail in Chapters ?? and ??.

1.3 NN Taxonomy

The wide of range of NNs discussed above can be classified into NNs with different topologies, different learning algorithms, and more. Figure ?? offers a classification which is apt to be only partial. The various names appearing in the boxes may not have been defined thus far, but they will be in Chapter ??. Nevertheless, we present the taxonomy at this point in order to bring some notion of order and familiarity to the NNs that have been discussed so far. One other preview that may be of interest to the reader whom at this point in the book is still deciding whether or not to buy this book is highlighted in the double-boxes of Figure ??; they represent the type of NN to which most of this book is devoted. In short, the type of NN that is almost entirely the focus of this book is the fully supervised, feedforward neural network. The reason for this is that this type of NN has the greatest affinity with traditional regression and classification - the primary topics of this book.

1.4 Exercises

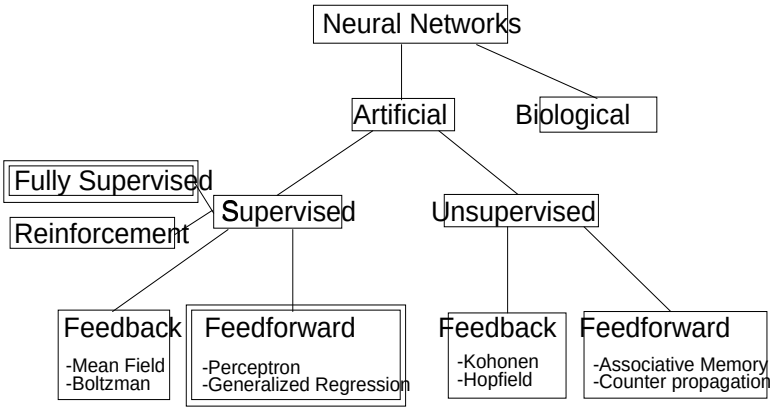


Fig. 1.1 The taxonomy of neural networks. The names written in smaller font in the boxes at the lowest end of the figure are examples of the corresponding type of NN.

Chapter 2
Multiple Regression

Suppose we have data on two variables, x and y. Several situations are possible: We may simply be interested in finding out if there is any relationship between x and y. On the other hand, it may be that there exists some theory that predicts a certain relationship between these variables, and we are interested in proving or disproving the theory. It may also be that one of these variables, say x, is always measurable, while y is only sometimes available for measurement, and the problem is to predict y from a knowledge of x. Or, perhaps, both variables are available over some range of values, but we are interested in the value of y, given a value of x that is outside the range of the available data. All of these problems fall under the general topic of regression.

Suppose x is the amount of television viewed per day and y is academic achievement. Never mind how these are measured; let us just assume that we have observations on these two quantities from 100 students. In the first situation, it is interesting in and of itself to find out if there is any relationship between television viewing and achievement level. In the second situation, we may discover that there exists no relationship between the two, in contrast to the linear relation predicted by some theory, or that a linear relationship between the two is inconsistent with a nonlinear relationship predicted by the theory.

Put a few more words about regression for social scientists....

2.1 Linear

Let us explore the case where the true underlying function is

$$y = \alpha_1 x + \alpha_0 + \epsilon, \tag{2.1}$$

which in terms of data translates to

$$y_i = \alpha_1 x_i + \alpha_0 + \epsilon_i. \tag{2.2}$$

Note that there is a different error term, ϵ_i, for every observation y_i. Conforming to tradition, Greek letters will represent true (or population) parameters, and Roman letters will represent the corresponding estimates (as obtained from a sample data set). These estimates can then be utilized to predict y from x, i.e.

$$y(x) = a_1 x + a_0. \tag{2.3}$$

For every observation, the difference between the actual value of y and the corresponding predicted value is called the residual, written as

$$e_i = y_i - y(x_i). \tag{2.4}$$

Note that α_0, α_1, and ϵ_i, are all unknown parameters. In fact, it is practically impossible to estimate all of these from a sample of size N, because i goes from 1 to N, and so there are $N + 2$ parameters to estimate from only N observations. That is why one typically estimates only α_0 and α_1, leaving the ϵ_i alone. As such, our predictions based on Eqn (??) will always differ from the observed values by an amount e_i. (Is this true?)

What criterion should be used for estimating the parameters α_i? We could try to minimize the sum of the residuals, since that is how much each observation differs from the predicted value. However, it is easy to show that the sum of the residuals is in fact identically zero (Exercise ??). (when is this not true?). In other words, on the average the prediction Eqn (??) overestimates just as much as it underestimates (is this true?). An alternative criterion would be to minimize the sum of the squares of the residuals.

List the assumptions that go into the analysis.
Point out that normality is an assumption that enters only in doing statistical tests.

2.2 Nonlinear

The "nonlinear" in nonlinear regression refers to nonlinearity in the parameters α, β, etc., not to the nonlinearity in the variable x. In other words, $y = \alpha log(x) + \beta$ is still a linear regression problem, because the relation is linear in α and β. Afterall, one can simply take the logarithm of the x values in the data and then perform a linear regression on the resulting values and y.

In fact, even many nonlinear regression problems can be transformed into linear regression problems. For example, even though the relation $y = log(\alpha x + \beta)$ is nonlinear in the parameters, rewriting it as $e^y = \alpha x + \beta$ suggests that the exponentiation of the dependent variable y reduces the problem to linear regression problem. Give a better explanation of this....

There are times, however, that no clever transformation of the data can reduce the nonlinear problem to a linear one. Give examples .....

2.3 Multiple, Linear

Illustrate interactions and multicollinearity.

2.4 Multiple, Nonlinear

Illustrate just how messy things can get.

2.5 Application: Television Viewing and High School Mathematics Achievement

This section examines the relationship between mathematics achievement and television viewing. The data consist of 13,542 high school seniors from the High School and Beyond project conducted by U.S. Department of Education, National Center for Education Statistics.

Data were obtained from a major longitudinal survey of American youth - High School and Beyond Series (1980-1986) - continued as the National Education Longitudinal Study, U.S. Department of Education, National Center for Education Statistics (1992). The data include information on 28,240 senior students, sampled by a two-stage stratified probability sample of 1,015 high schools. Schools were selected with a probability proportional to their estimated enrollment, and within each school data were obtained on 36 seniors chosen at random. Subjects with multiple punches or missing data were excluded, leaving a sample of 13,542 seniors for the analysis.

Based on prior literature, the following variables were selected to examine the underlying process: The dependent variable was taken to be mathematics performance, the independent variables were amount of television viewing, gender, ability, and ethnicity (i.e., viewer attributes), socio-economic status, father's occupation, father's education, mother's occupation, mother's education (i.e., parental background), parents' knowledge of what the student does, father's monitoring of school work, and mother's monitoring of school work (i.e., parental involvement), reading for pleasure, and visiting friends (i.e., leisure activities). The variables that require a description are: Mathematics performance is measured in terms of the average of Mathematics I, and Mathematics II standardized tests. The range of the mathematics composite t-score is 27 to 72, with a mean of 50, and a standard deviation of 10.

Television viewing time is gauged by the question, "During weekdays, about how many hours per day do you watch TV?" Responses range from "don't watch TV during day" (coded 0) to "five or more per day" (coded 6). Intermediate responses are "less than 1 hour," "1 to 2 hours," "2 to 3 hours," "3 to 4 hours," and "4 to 5 hours."

Ability is a composite score of six tests measuring students' abilities - picture-number, mosaic comparisons (I and II), visualization in three dimensions, and vocabulary (I and II). This measure of ability has been put forth by Keith, et al. (1986). Each score is standardized (t-score; mean = 50, standard deviation = 10), and averaged. The ability variable ranges from 28 to 71.

Socio-economic status (SES), as defined in the High School and Beyond Project, is a composite score of father's occupational status, mother's and father's educational attainment, family income, and possessions in the home (whether the family had a daily newspaper, an encyclopedia, a typewriter, two or more cars, more than 50 books, a room of the student's own, and a pocket calculator). Each question is converted to a z-score, and then averaged. Based on this standardized score, each student is classified as belonging to a low, middle, or high SES, depending on whether he/she belongs to the lowest quartile, middle two quartiles, or the highest quartile in the distribution.

Here, put the results of first-order linear regression
second-order linear regression
maybe third-order linear regression
and one nonlinear regression.

2.6 Exercises

1) Show that the sum of the residuals in linear regression is identically zero.

Chapter 3
Classification

This goes mostly like the regression problem. The (big) difference is that the dependent variable is now discrete. This is one valid way of looking at a classification problem. In fact, it is possible to perform a classification with a regression model, e.g. logistic regression....

3.1 Parametric

General stuff.

3.1.1 *Discriminant Analysis*

Details of Discriminant Analysis (DA) can be found in (McLachlan, 1992). In its simplest form, the data is assumed to be multivariate normal (section??), and classification is made based on whether or not an observation is less than or greater than some threshold value according to the posterior probability of each class.

Specifically, the Likelihood of an observation, x, given that it belongs to the i^{th} class, is assumed to be

$$L_i(x) \sim \frac{1}{\sqrt{det\ V_i}}\ exp^{-\frac{1}{2}(x-\mu_i)^T\ V_i^{-1}\ (x-\mu_i)}\ , \tag{3.1}$$

where μ_i is the vector of the means and V_i is the covariance matrix for the i^{th} class (here $i = 0, 1$), all estimated from a training data set. Then, an observation, x, is classified into the class with the larger posterior probability, $P_i(x)$, which in turn is derived from Bayes' theorem:

$$P_i(x) = \frac{p_i L_i(x)}{p_0 L_0(x) + p_1 L_1(x)}, \tag{3.2}$$

where p_0, p_1 are the prior probabilities for the two groups, discussed below. Of course, $p_0 + p_1 = 1$ and $P_0 + P_1 = 1$.

It is easy to show (McLachlan, 1992) that the decision criterion (or the discrimination function, $\log(P_1(x)/P_0(x))$), is quadratic in the quantity x; such an analysis is referred to as a Quadratic Discriminant Analysis (QDA). However, if $V_0 = V_1$, then the decision criterion is linear in x, leading to a Linear Discriminant Analysis (LDA). The main advantage of the latter is in allowing for the interpretation of the linear coefficients as "predictive strengths" of the corresponding variables.

There are some problems with this business of interpreting the coefficients as measures of predictive strengths. Collinearity...

In the one-variable case, things are pretty simple:

$$\log(P_1(x)/P_0(x)) = \frac{1}{2}D^2(x),$$

where the so-called discriminant function, $D^2(x)$, is given by

$$D^2(x) = (\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2})x^2 - 2(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2})x + (\frac{\mu_0^2}{\sigma_0^2} - \frac{\mu_1^2}{\sigma_1^2}) + 2\log(\frac{\sigma_0}{\sigma_1}) - 2\log(\frac{1-p_1}{p_1}). \quad (3.3)$$

The means and the standard deviations are estimated from the sample data, and then an observation, x (either from the same data or an independent data), is assigned to group 1 if $D^2(x) > 0$ (i.e., $P_1(x) > P_0(x)$), otherwise it is classified (forecast) as a 0 (i.e., $P_0(x) > P_1(x)$). An x that yields $D^2(x) = 0$ can always be assigned to one of the groups on a random basis.

Typically, one estimates the parameters from one sample and then obtains a measure of performance by computing the measure for an independent sample; the former sample is called the training set, and the latter is called the validation (or test) set. (Later, in dealing with NNs, we shall find reason for distinguishing further between the validation and the test set.)

In the fortunate situation where $\sigma_0 = \sigma_1 = \sigma$, (i.e., if the data is so-called homoelastic) then the discriminant function becomes linear in x. This has great utility in the multivariate case, because then the coefficients of the various x-terms would represent the predictive strength of the respective independent variables, if certain requirements are met (see below). This is the so-called Parametric Linear Discriminant Analysis (PLDA).

Discuss the traps in interpreting the coefficients as measures of predictive strength.

In summary, PLDA is a popular classification method (Huberty, 1994; McLachlan, 1992) with several explicit assumptions; the data are assumed to be multivariate gaussian (normal), and the different classes are assumed to have equal covariance matrices (homoelastic).

3.1.2 *Unbalanced Classes*

Show the effect of a priori probabilities....

3.2 Non–parametric

Discuss some non-DA models....

Show that kernels can be employed to relax the assumptions of DA.....

3.3 Application: Predicting Television Extreme Viewers and Nonviewers

It is instructive to go through a practical example from beginning to end. In this section, we will illustrate the application of PLDA to the problem of discriminating between extreme television viewers and nonviewers.

The data for this study is taken from the 1988, 1989, and 1990 General Social Surveys (GSS), conducted by the National Opinion Research Center (NORC), using a stratified, multistage area probability sample of clusters of households in the continental United States (Davis & Smith, 1990). Respondents were randomly selected adults age 18 and over, one from each household.

In this example, the definition of non and extreme viewers is as follows (Hirsch, 1980): Nonviewers are those who watch zero hours of television and extreme viewers are those who watch television 8 or more hours per day. These groups comprise the bottom and top 4% of the viewing hours distribution, which are one standard deviation below and above the mean of 2.9 hours for the entire NORC General Social Survey data set.

Three experiments can be performed each with a different set of variables as inputs. The input variables for the three experiments are: Experiment 1 - Demographic variables: age, gender, ethnicity, education, family income, job classifications, size of the population, and region of interview. Experiment 2 - Family-related variables: number of household members, marital status, number of members under 6 years old, number of members from 6 to 12 years old, number of members from 13 to 17 years old, number of members over 17 years old, and general happiness. Experiment 3 - lifestyle/social activity-related variables: strength of religion, frequency of attending religious services, frequency of spending social evening with relatives, frequency of spending social evening with someone who lives in one's neighborhood, frequency of spending social evening with friends who live outside the neighborhood, frequency of going to a bar or tavern, frequency of spending social evening at one's parents, frequency of spending social evening with relatives, frequency of spending social evening with someone who lives in one's neighborhood, frequency of spending social evening with friends who live outside the neighborhood, frequency of going to a bar or tavern, frequency of spending a social evening at one's parents, frequency of spending a social evening with a brother or sister, Amount of hours listening to radio per day, and frequency of reading the newspaper.

One may be tempted to take the training set to be the 1988 and 1989 surveys,

and the validation set from the 1990 survey. We shall illustrate this particular set-up, below; however, as we shall see later (in section ??) the better practice is to randomly select many training and validation sets from the entire data set. This is called bootstrapping ....

Those who answered zero hours to GSS question, "On the average day, about how many hours do you personally watch television?", were classified as nonviewers and those who answered 8 or more hours were classified as extreme viewers. Based on this definition, 65 (32 from 1988 and 33 from 1989) nonviewers are selected for training, and 29 (from 1990) for validation. As for extreme viewers, 100 (53 from 1988 and 47 from 1989) respondents are selected for training and 33 (from 1990) for validation. In this way, we use data from 1988 and 1989 for training, and the 1990 data for prediction/validation. From this subsample data set, data having more than three missing pieces of information were excluded, and one or two missing responses were replaced by their respective means. This brought the number of nonviewers to 49 and extreme viewers to 66 for the training set, and 27 nonviewers and 24 extreme viewers for the validation set.

The variables included in the analysis are demographic, family-related, and lifestyle/social activity-related. The content of these three classes of variables are listed in Appendix A, as appearing in GSS, and we have labeled them numerically, as (1), (2), and (3), respectively. Treating each class as a separate "variable", we adopted a stepwise approach in identifying the viewer characteristics in terms of these classes. Then, in an obvious notation, the 7 different cases are written as: (123), (12), (13), (23), (1), (2), and (3). For example, case (12) refers to the combined set of demographic and family-related variables, and (123) refers to set of all 25 variables.

The statistical package SAS^{TM} was employed for performing the PLDA part of our analysis. Data were presented to the PROC DISCRIM algorithm, which is designed to calculate the vector of mean scores, as well as the pooled and within-group dispersion matrices for LDA and QDA, respectively. This is the training phase of the analysis.

The algorithm then returns to Equation ??, with the mean scores and the dispersion matrix as determined from the training phase, and substitutes the individual scores for each case into x. The resulting value of the function $C(x)$, is then used to reclassify each case into the two populations.

The validation phase consists of substituting new values of x into Equation ?? and then classifying, using the same mean scores and the dispersion matrix as obtained from the training phase. The number of correct and incorrect classifications of the validation is a better measure of DA's goodness-of-fit, since it tests the model's predictive capabilities. Due to its simplicity, an PLDA is performed first, followed by a QDA.

Put Results here.

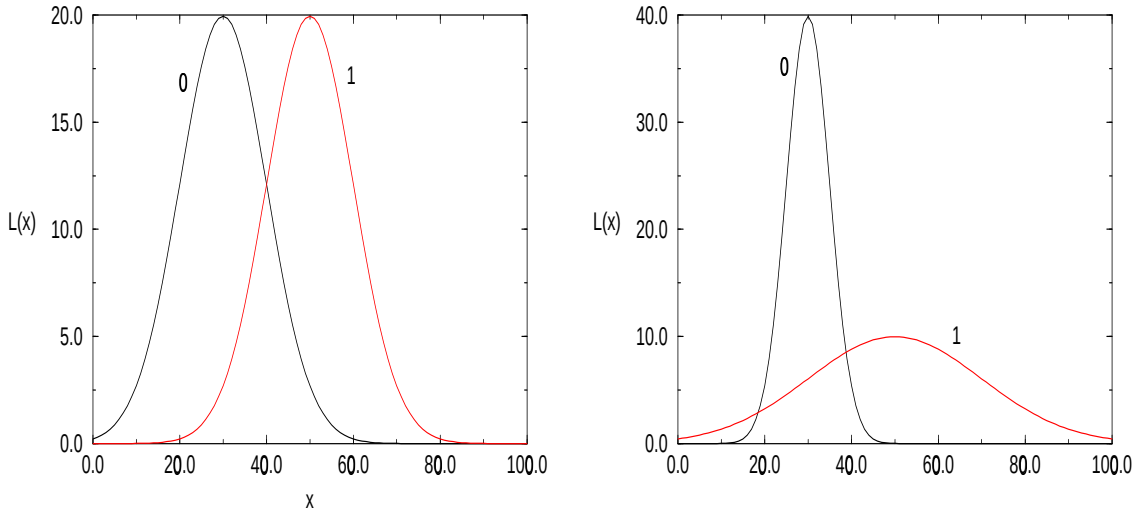


Fig. 3.1 Two normal distribution curves with means at $\mu = 30$ and $\mu = 50$, with equal variances (left) and with unequal variances (right).

3.4 What Does the (Multivariate) Normal Distribution Look Like?

Clearly, obtaining an estimate for the underlying function does not call for any assumptions regarding the distribution of the data. However, to gain any measure of statistical significance or the goodness-of-fit does require some further assumptions regarding the distribution of the data. Usually, the distribution is assumed to be the multivariate normal (Gaussian) distribution (refs??).

The case of a single variable is almost needless to discuss. The equation for the (univariate) normal distribution is

$$y = \frac{1}{\sqrt{2\pi}\sigma}\, exp^{-\frac{(x-\mu)^2}{2\sigma^2}} , \tag{3.4}$$

but it is equally valid and useful to simply plot y as a function of x for some values of the mean and the variance. For example, one look at Figures ??a and ??b reveals the appearance of the normal distribution and its dependence on its two parameters - the mean, μ, and the variance, σ^2.

Probably all readers already know that a multivariate normal distribution basically looks like some bell-shaped surface when plotted against the arguments x_1 and x_2. However, it is much less obvious how the covariance matrix affects the shape of the distribution. To simplify the task, we shall consider an example involving two variables, x_1 and x_2. The equation for such a distribution is

$$\frac{1}{\sqrt{(2\pi)^2 |\det V|}}\, exp^{-\frac{1}{2}(\vec{x}-\vec{\mu})V^{-1}(\vec{x}-\vec{\mu})} , \tag{3.5}$$

where $\vec{x}$ represents the vector (x_1, x_2) and $\vec{\mu}$ represents the vector of the means of the two variables (μ_1, μ_2). The covariance matrix can be written out explicitly as

$$V = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}. \tag{3.6}$$

We may already anticipate that the peak of the bell will occur at the point with coordinates $(x_1, x_2) = (\mu_1, \mu_2)$, and indeed, this can be seen in Figure ?? wherein the two means are set to zero. The graphs in Figure ?? enumerate the following cases:

(1) When the two variables are uncorrelated and have equal variance. In equations, $\sigma_{12} = 0$ and $\sigma_1 = \sigma_2 = 2$.

(2) When the two variables are uncorrelated ($\sigma_{12} = 0$), but have different variances ($\sigma_1 = 2, \sigma_2 = 4$).

(3) When the two variables are correlated ($\sigma_{12} = 1$) and have equal variance ($\sigma_1 = \sigma_2 = 2$).

(4) When the two variables are correlated ($\sigma_{12} = 1$) and have unequal variance ($\sigma_1 = 2, \sigma_2 = 4$).

Clearly, the last case is the most general, but the first three cases are instructive to examine, as well.

Evidently, in the first case, the distribution is perfectly bell-shaped and has a circular cross-section. In the second, the bell is stretched in one direction and has an elliptic cross-section; the stretching is in the direction of the variable with the larger variance. In the third case, a nonzero correlation causes the bell to rotate so that the major axis of the ellipse is along some direction other than x_1 or x_2. From the fourth case we see that this rotation occurs even when the variances are unequal. In short, the covariance term σ_{xy} has the effect of rotating the ellipse. Say more on this ...
Better pictures ??

3.5 Decision Boundaries

Given the discussion of section?? regarding the shape of the multivariate normal distribution, it is possible to say a few things about what types of decisions are most suitable for DA. This is the topic of decision boundaries.

The term "boundary" may suggest that we are talking about some regions separated by some curve. In fact, a boundary can be a point, a curve, a surface,

or a hypersurface, etc. It all depends on the number of classes and independent variables, but not too simply.

For two classes and one (continuous) independent variable, the situation could be as depicted in Figure ??. Here, there are two regions (lines) and the boundary is a single point. It may also be that two points are required to separate the classes. Then the boundary consists of two points.

The situation can be viewed in terms of the probability density curves for the two classes (Figure ??). One figure suggests that one point is sufficient to separate the two classes, namely the value of the x-axis at which the two curves cross. However, when the two curves have different variances, then the two curves cross at two points! In other words, two decision boundaries (or thresholds) are required. The first case is an example of a linear boundary while the second case represents a nonlinear boundary. Explain better and more ...

The terminology can be further clarified by considering the case of two classes and two independent variables. Figure ?? shows a linear boundary between two classes, while Figure ?? shows an example of a nonlinear boundary.

PLDA, in fact, has linear boundaries.....

PQDA has elliptic, parabolic, hyperbolic, etc. boundaries, ...
(but maybe we don't want to get into so much detail).

Give an example of a Kernel-based DA....

3.6 Exercises

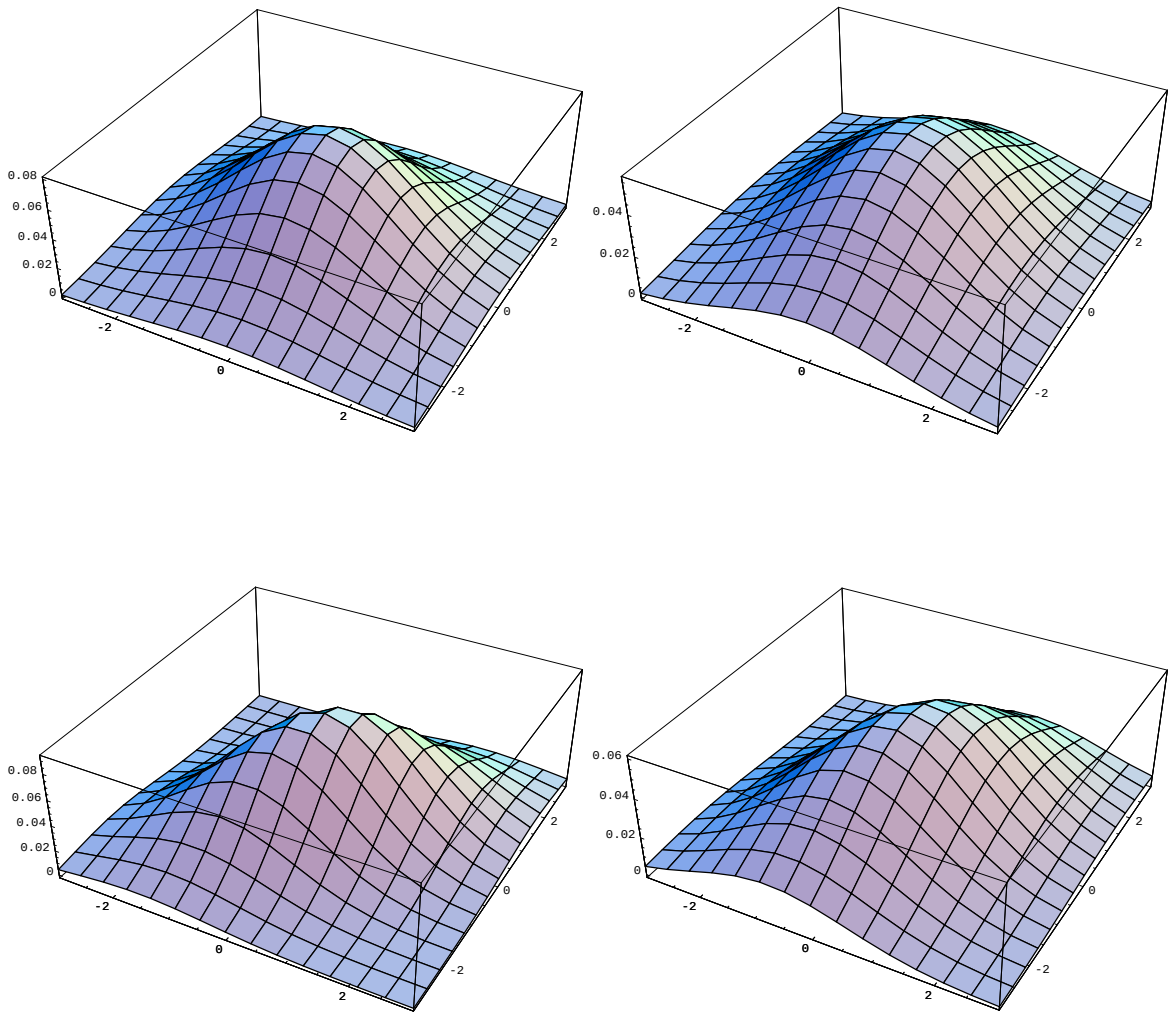


Fig. 3.2 Bi-variate normal distributions for two variables that are uncorrelated and equivariant (top, left), uncorrelated but with different variances (top, right), correlated with equal variances (bottom, left), and correlated with different variances (bottom, right).

Chapter 4
Measures of Performance

The question of what exactly is a proper measure of performance depends on a number of things one of which is the nature of the problem itself....

The statistics literature on that subject is extensive (refs)....

In this section, we shall point out some of the "bad" practices and some "good" alternatives .....

A frequently neglected fact is the multi-faceted nature of performance. For example, everyday we hear about the rise or fall of unemployment. But what is a measure of unemployment? We know it's a number, but the number of what? The number of high-pay professional jobs, or that of minimum-pay jobs? ....

Another example is found in the world of economics. We hear phrases like "the economy is up," but what precisely is that "economy?" Well, this one is a particularly thorny question because economic systems are typically highly complex, and because it deals with money! That's why there are many "indeces" (or "indexes") of economic status, each measures a different facet of the problem....

Another place where the multi-faceted nature of performance plays a role is in statistics itself and is referred to as *model comparison.* Basically, the question there is whether my model is better than yours, or whether the model I constructed today is better than the one I came-up with yesterday. Naturally, it all depends on what we mean by "better," but the problem is that few people realize how difficult it can be to define "better." Meanwhile, a great many decisions are made on a daily basis with complete disregard of the complexity of the issue; individuals are hired, or fired, and social policies are implemented, or discontinued, all based on the relative performance of the individuals or the social policies in terms of a single measure.

The fact is that any comparison between the performance of one model and another model in terms of a scalar (one-dimensional) quantity is apt to be incomplete. Consider a simple classification problem involving two classes. The performance of such a classifier is best expressed in terms of a 2×2 contingency table (below). But even this contingency table, representing the performance of the binary classification of two classes (of fixed sample size) has two degrees of freedom. Therefore, in a binary classification task a faithful comparison would require at least two inde-

pendent measures of classification performance.

It is possible that one model is better than another in terms of all the relevant measures of performance, but that situation is neither garaunteed nor likely. However, there are times that a scalar measure must be employed. Example ....

Two types of scalar performance measures must be distinguished - continuous and discrete. The former are computed from continuous quantities, like the dependent variable in a regression problem, while the latter are computed from a contingency table. Of course, one may reduce the continuous quantity into a binary one by placing a (decision) threshold on it and then forming the contingency table. Mean-square error and cross-entropy are examples of the former, and percent correct is an example of the latter....

A proper choice of the measure is especially important because it is entirely possible that method A will outperform method B in terms of one measure of performance, but not in terms of another. In many applications, however, the quantity that is minimized is the mean-square error even though that choice is justified only if the probability density of the dependent variable is gaussian (or at least continuous, or bell- shaped), for only then will the parameter estimates coincide with the maximum likelihood estimates. However, such models are often unjustifiably employed for classification problems where the dependent variable is discrete, e.g., binary, 3-valued, etc. (refs??).

One other common practice is to assess performance from the same data set (training set) that is used for estimating the parameters of the model (refs??). The problem is that the performance of all parametric models on the training set is positively biased. An independent data set is required for an unbiased assessment.

Since we are speaking of data, it must be emphasized that it is entirely possible that method A will outperform method B on one data set, but not on another. This contingency is one that requires only a confession, in that it is sufficient to acknowledge that any empirically established superiority of one method over another is specific only to the particular data set (and measure) being analyzed.

The moral of this section is that performance has many components, each of which can be gauged by some measure. One should employ all the relevant measures of performance independently, if one can, rather than combining them into a single, scalar measure. ....

4.1 Regression

Some measures of performance appropriate for regression problems are

$$\text{Mean Absolute Error} \quad = \quad \frac{1}{N}\sum_{i=1}^{N}|y_i - t_i|, \tag{4.1}$$

$$\text{Mean Square Error} = \frac{1}{N}\sum_{i=1}^{N}(y_i - t_i)^2, \tag{4.2}$$

$$\text{Cross-entropy} = -\sum_{i=1}^{N}(t_i \log \frac{y_i}{t_i} + (1-t_i)\log \frac{1-y_i}{1-t_i}). \tag{4.3}$$

Each is appropriate for a given distribution of the data. Mean absolute error is appropriate for a Laplacian distribution, and mean square error assumes a normal distribution, while cross-entropy is appropriate for classification problems where the dependent variable is continuous but the target is discrete. Yes, it is true, a regression model may be employed for classification purposes - example, logistic regression (or NNs, section??).

4.2 Application: ....

Another application.

4.3 Classification

Some measures of classification performance are ill-behaved or inappropriate in certain situations. For instance, the use of the commonly employed measure "percent correct" is misleading if the classes are not equally represented in the data set. This is so because that measure does not take into account chance or random guessing, and as a result, even a random classifier can yield a 99.9% accuracy if for instance one class is much more frequent that another (Paik, 1998). Of course, this shortcoming is readily exposed if the statistical significance of the measure is considered, but often it is not.

It is important to examine the behavior of these measures in certain special situations, such as deviations from normality, or small sample sizes. Such matters have been considered by Hammond and Lienert (1995), and by Parshall and Kromrey (1996). Discuss these in some detail.....

Another special, yet ubiquitous, situation arises when the class sample sizes are disproportionate. Many measures of performance have been examined in this limit by Marzban (1998), wherein it has been shown that many measures are ill-behaved (or biased) in that they over-estimate the true performance of the classifier. Discuss this, too ...

Give an exhaustive list of all measure commonly employed in the social sciences for discrete measures.....
for continuous measures ....

Specialize to the 2×2 case, for illustrative purposes....

Of all the discrete measures listed above, three that appear to be relatively

"healthy" are Pearson's χ^2, the mean-square contingency, ϕ, and the likelihood ratio chi-squared, LR, defined as

$$\chi^2 = \sum_{i=1}^{4} \frac{(C_i - E_i)^2}{E_i^2}, \tag{4.4}$$

$$\phi = \sqrt{\frac{\chi^2}{N}}, \tag{4.5}$$

$$LR = \sum_{i=1}^{4} C_i \log(\frac{C_i}{E_i}). \tag{4.6}$$

where C_i are the elements of the 2×2 contingency table

$$\text{C-table} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \#\text{ of 0's predicted as 0} & \#\text{ of 0's predicted as 1} \\ \#\text{ of 1's predicted as 0} & \#\text{ of 1's predicted as 1} \end{pmatrix} \tag{4.7}$$

$$= \begin{pmatrix} . & \text{false alarms} \\ \text{misses} & \text{hits} \end{pmatrix}. \tag{4.8}$$

and E_i are the elements of the expected matrix, i.e., the contingency table that would ensue upon random guessing:

$$E = \frac{1}{N} \begin{pmatrix} (a+b)(a+c) & (a+b)(b+d) \\ (c+d)(a+c) & (c+d)(b+d) \end{pmatrix}. \tag{4.9}$$

The total number of 0's is given by $N_0 = a + b$, that of 1's is $N_1 = c + d$, and the total sample size is $N = N_0 + N_1$. As mentioned previously, this table has only 2 degrees of freedom; a general 2×2 matrix has 4 degrees of freedom, but with the 2 constraints $N_0 = a + b$ and $N_1 = c + d$, that number is reduced to 2.

It is wise to normalize all the measures in Eqn (??) so that a perfect classification of both classes (i.e., a diagonal contingency table) will yield a value of 1, while random classification (i.e., C = E) will yield a value of 0.

4.4 Application: High School Delinquency

In this section an example involving unbalanced classes will be studied in order to illustrate some of the aforementioned defects of the measures.

The data are taken from the first follow-up (1990) of the National Education Longitudinal Study (NELS), (U.S. Department of Education, National Center for Education Statistics 1992), base year 1988. The 1990 student component collected basic background information about students' school and home environments, participation in classes and extracurricular activities, current jobs, and students' goals, aspirations, and opinions about themselves. This component also measures 10th grade achievement and cognitive growth between 1988 and 1990 in the subject areas of mathematics, science, reading, and social studies. The 20,706 subjects were

all 10th grade students in the United States during the 1989-1990 school year. The sampling was done in a two-stage sampling process, distributed across 1,500 schools, involving the selection of a core group of students who were in the 8th grade sample in 1988. Based on prior literature (Evans, Cullen, Burton, Dunaway, Payne, and Kethineni 1996; Kendall-Tackett 1996; Simons, Whitbeck, Conger, and Conger 1991; Watts and Wright 1990), 78 variables were selected as the independent variables (Appendix B), and the dependent variable was in-school suspension.

Some amount of preprocessing of the data is almost always necessary, and even beneficial, before any analysis. Here, all observations (students) with any missing data are neglected; all the independent variables are standardized (i.e. transformed into z-scores; and the dependent variable - the number of in-school suspensions is dichotomized into 0 and 1, for the nonsuspended and suspended, respectively.

After the preprocessing, the remaining 18,075 cases were randomly partitioned into a training set (12,000) and a validation set (6,075), four times, for cross-validation. The mean and the 90% confidence intervals of the validation performance measure, S, over the four different sets were then computed. Finally, the three measures of classification performance - χ^2, ϕ, and LR - were also computed; since these measures are discrete, a decision threshold was placed on the outputs. The threshold was varied in 0.01 increments and the validation performance measures were computed at each increment. In this way one can identify the optimal value of the decision threshold and the corresponding value of the performance measure.

Often the two classes are artificially equalized by including an equal number of 1s and 0s in the training set (but not in the validation set). This balancing of the classes is also believed to enhance the performance of an NN (Masters, 1993). It is important to emphasize that changing the class sizes in the training set robs the output from being interpreted as a posterior probability; that interpretation is valid only if the classes in the training set are represented according to their a priori probabilities.

The Pearson correlation coefficients, r, between the 78 independent variables and the dependent variable are plotted in Figure ??. The height of each bar in the graph is a measure of the linear correlation, between the independent variables and the dependent variable. The utility of this figure is in allowing for the selection of the input variables that are most correlated with in-school suspension.

Here, put results of
PLDA
PQDA
and maybe PCA

4.5 Exercises

Chapter 5
Neural Networks

As we have seen, a NN refers to a collection of some number of entities, interacting with each other. The entities are called nodes, and they may be observed attributes or variables, or simply abstract variables that do not represent any physical quantity.

As discussed in section??, the most general architecture would appear as that shown in Figure ??, where every node is connected to, and is interacting with, all other nodes. This type of NN is called a Feed-back NN, because the interaction between any two nodes is bi-directional. Such networks have been employed in the simulation of memory (Hebb, 1949; Hopfield, 1982), optimization problems (Hopfield & Tank, 1985), and in data analysis techniques such as feature recognition (Kohonen, 1984).

NNs come in a variety of topologies. An exact definition would take us to technical ends. Instead let us explore that topic with examples and figures. This will alow the reader to "infer" a definition. Figure ?? is an example of a fully connected NN - and one can see why it is called that. One may decide, for reasons to be discussed later, to arrange the nodes on layers. The simplest case of two layers defines a *perceptron*, while many layers constitute a multilayered perceptron. These are two different topologies. Finally, one may allow for feed-back, leading to what is called a recurrent NN. All of these will be discussed in further detail in the next chapter; however, they all share some basic structure which we can set-up at this point.

Such NNs have two variants: those with symmetric weights and those with the more general asymmetric weights. In this network, the various features (patterns) are stored as dynamical equilibria of the system. One begins with the energy function

$$E = -\frac{1}{2}\sum_{i\neq j}\sum_{j}\omega_{ij}\sigma_i\sigma_j \quad , \tag{5.1}$$

where the numbers ω_{ij} represent the strength of the weight between the nodes x_i and x_j. It can then be shown (Müller & Reinhardt, 1991) that the various minima of this energy correspond to stable equilibrium states, where the weights insure the

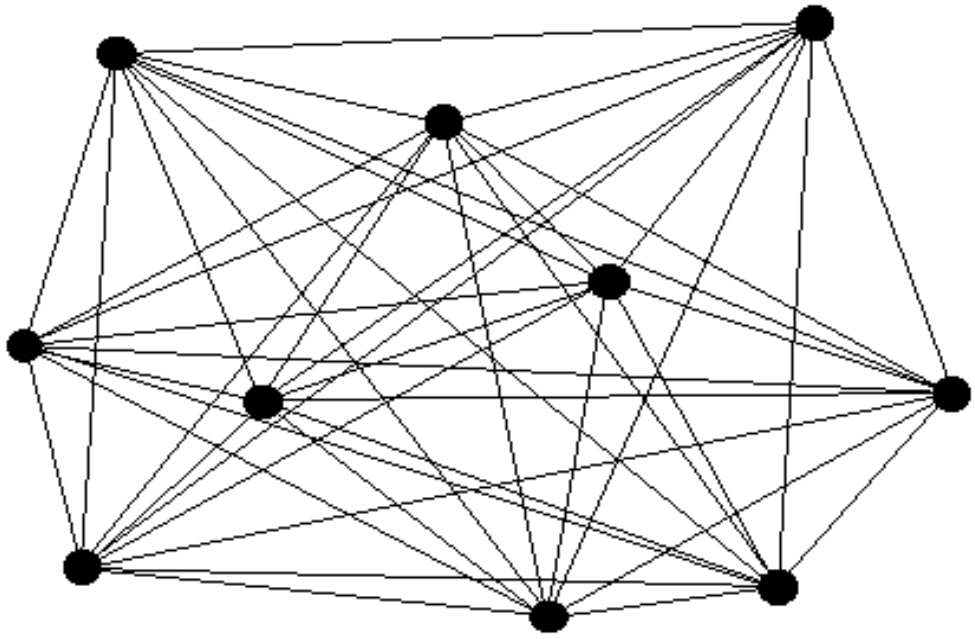


Fig. 5.1 A fully connected neural network.

correct pattern to emerge among the various nodes. For instance, the letter "A" can be considered as a collection of nodes in a square array, some of which are 0, while others (those composing the letter itself) are 1. The set of weights necessary for the storage of this letter are those that minimize the energy function in equation (??).

As for the asymmetric NN (i.e. that with asymmetric weights), such NNs fall in the category of non-equilibrium systems - a far more difficult problem to study at the practical and the theoretical level. It is well-known (Müller & Reinhardt, 1991) that such systems do not have a unique equilibrium, and tend to move between different equilibria. For example, such temporal sequences of memories have been employed to explain the course of events and actions, which otherwise would consist of meaningless single pictures.

Another type of NN incorporates a given architecture (displayed in Figure ??), wherein non-interacting (independent) nodes are placed in layers, interacting with the nodes in other layers, and only in one direction, hence the name Feed-forward NN (??). These constitute a special class of multilayered perceptrons. Their utility is in finding a function that relates the outputs of the network to the inputs.

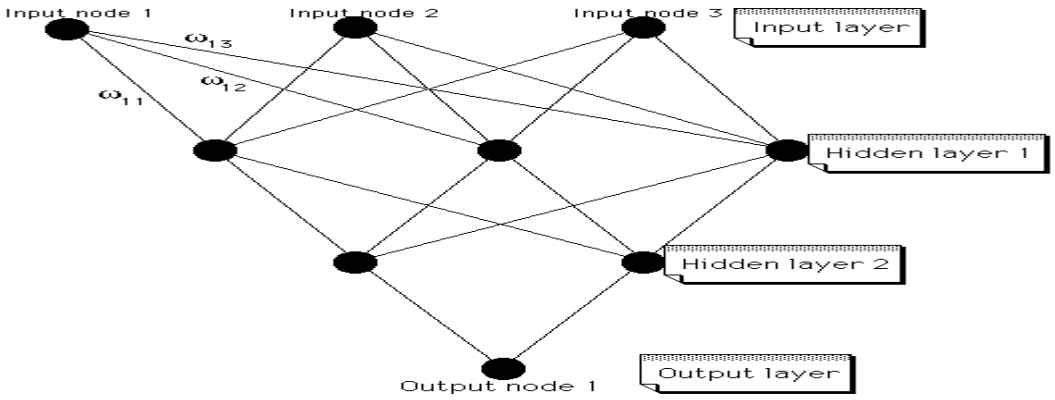


Fig. 5.2 A neural network with two hidden layers.

Here, the input nodes and the output nodes correspond to the independent variables and the dependent variables of the function, respectively. The hidden nodes have no physical interpretation apart from allowing the network to deal with nonlinearities in the data. It has been shown that there are some functions (models) that cannot be dealt with by a network with no hidden layers. However, it has also been shown that a network with two hidden layers is sufficient for learning *any* function (Figueiredo, 1980; Hecht-Nielsen, 1987); see next chapter.

The utility of the former type of network is in its ability to simulate the interactions among all the variables in a more assumption-free manner than the layered NN. One may "physically" alter the value of any given node and consequently observe the behavior of all other nodes in the network. Given the lack of a unique

stable equilibrium, a change in the value of one variable may cause the network - e.g. brain, society - to go through a phase transition from one equilibrium state to another.

....

Multilayered percpetrons are often presented as a novel statistical method with no assumptions. However, recently, it has become evident that they are by no means assumption-free, although the assumptions may be considered "milder" and more implicit than those of many other methods (Bishop, 1996; Masters, 1993; Sarle, 1994a; Ripley, 1996). This flexibility has given rise to an extravagant use of such NNs, even in situations in which some of the implicit assumptions are violated.

One of the justifications for such practices is based on their capability to model highly nonlinear relationships and nontrivial interactions among the variables of the model. It is often said that this property implies that such NNs can outperform all other methods. However, for a given data set, such an NN may be outperformed by a method with many restrictive assumptions. For example, if the model underlying the data is a polynomial, then polynomial regression will outperform a multilayered perceptron. In other words, it is not true that the milder and more implicit assumptions of this type of NN automatically renders it superior to the alternatives.

5.1 Artificial Intelligence (AI)

Within AI techniques, there exists another branch that goes by the name of expert systems. It is important to distinguish NNs from expert systems. The latter are "rule-based" which simply means that the operator must have a priori knowledge of the model to be simulated. These algorithms are a set of "if ..., then ..." statements that result in some outcome. The a priori knowledge necessary for the implementation of such a methodology is often nonexistent, directly leading to the invention of models. Models, in this traditional sense, provide one with a framework which can be employed to derive the rules governing the system, and the rules are then tested on observational data. Of course, the laws of a system may be derived from inspiration! An example is Einstein's theory of General Relativity, where the theory and the rules were in no way derived from (although, were later tested by) observational data. This, however, is a most singular example in the history of science.

NNs are most useful precisely when there is no a priori knowledge of the underlying function....

....

5.2 Learning Paradigms

As seen, there exists a wide variety of NNs for performing an equally wide range of tasks, but the way in which NNs are trained can generally be divided into two

paradigms - unsupervised and supervised - with each composed of several more specialized schemes.. The key idea in the former is self-organization, in that such an NN is designed to automatically search for salient features in the data. Such NNs are analogs of the traditional methods for cluster analysis. Supervised NNs, on the other hand, are analogous to regression and classification methods wherein the dependent variable is known and is utilized in the training procedure. Of course, cluster analysis can be employed for classification purposes as well; however, as we shall see, the classification capabilities of supervised NN are far superior to those of unsupervised NNs.

....

5.2.1 *Supervised*

A particular type of supervised NN is the so-called Multilayered Perceptron wherein the network has a layered architecture with the nodes on a given layer interacting only with the nodes on the adjacent layers. The input layer contains the nodes that represent the independent variables, and the nodes of the output layer represent the dependent variables of the problem. The existence of hidden layers simply renders the NN nonlinear. Usually, the hidden nodes have no physical meaning, though there are instances in which they can be interpreted as compressed representations of the data. An NN with one hidden layer containing H hidden nodes can be written as a single parametrization:

$$f\left(\sum_{i=1}^{H}\omega_i\, f(\sum_{j=1}^{N_{in}}\omega_{ij}x_j-\theta_j)-\theta\right). \tag{5.2}$$

where the ω's and θ's are all weights (parameters) to be estimated, the x_j are the N_{in} input nodes, and H is the number hidden neurons (on a single layer).

The function $f(x)$ is the so-called activation function. Clearly, if $f(x)$ is linear in x, then Eq. (??) is linear in the parameters, and the NN can be interpreted as a linear regression. If $f(x)$ is the logistic function, defined as

$$\frac{1}{1+exp^{-x}}, \tag{5.3}$$

then the NN can represent a logistic regression model (more on this, below). There is a large family of activation functions that allow the NN to perform different tasks, but for classification problems (as opposed to regression problems) it is sufficient to use the logistic function (refs??).

To get an idea of what the logistic function looks like, Figure ?? shows the behavior of the function

$$y=\frac{1}{1+exp^{-(ax+b)}}, \tag{5.4}$$

for a range of the values a and b. First, note that whereas x can take any value, y is contained in the range 0 to 1. Explain the consequences of this...

Furthermore, the value of a determines how quickly the curve rises and b shifts the curve to the left (positive b) or to the right (negative b). Also note that a large value of a (e.g. 4) leads to a highly nonlinear curve whereas for small values of a (e.g. 0.5) the function is almost linear. We shall return to this point when we discuss overfitting in Chapter ??.

5.2.2 *Unsupervised*

Explain autoassociative memory.

5.2.3 *Reinforcement*

In a fully-supervised NN, the target value (i.e. the answer one desires at the output) is propagated and involved in the training phase....

In Reinforcement learning the only quantity that is involved in training is how well the NN is generally performing....

5.3 Exercises

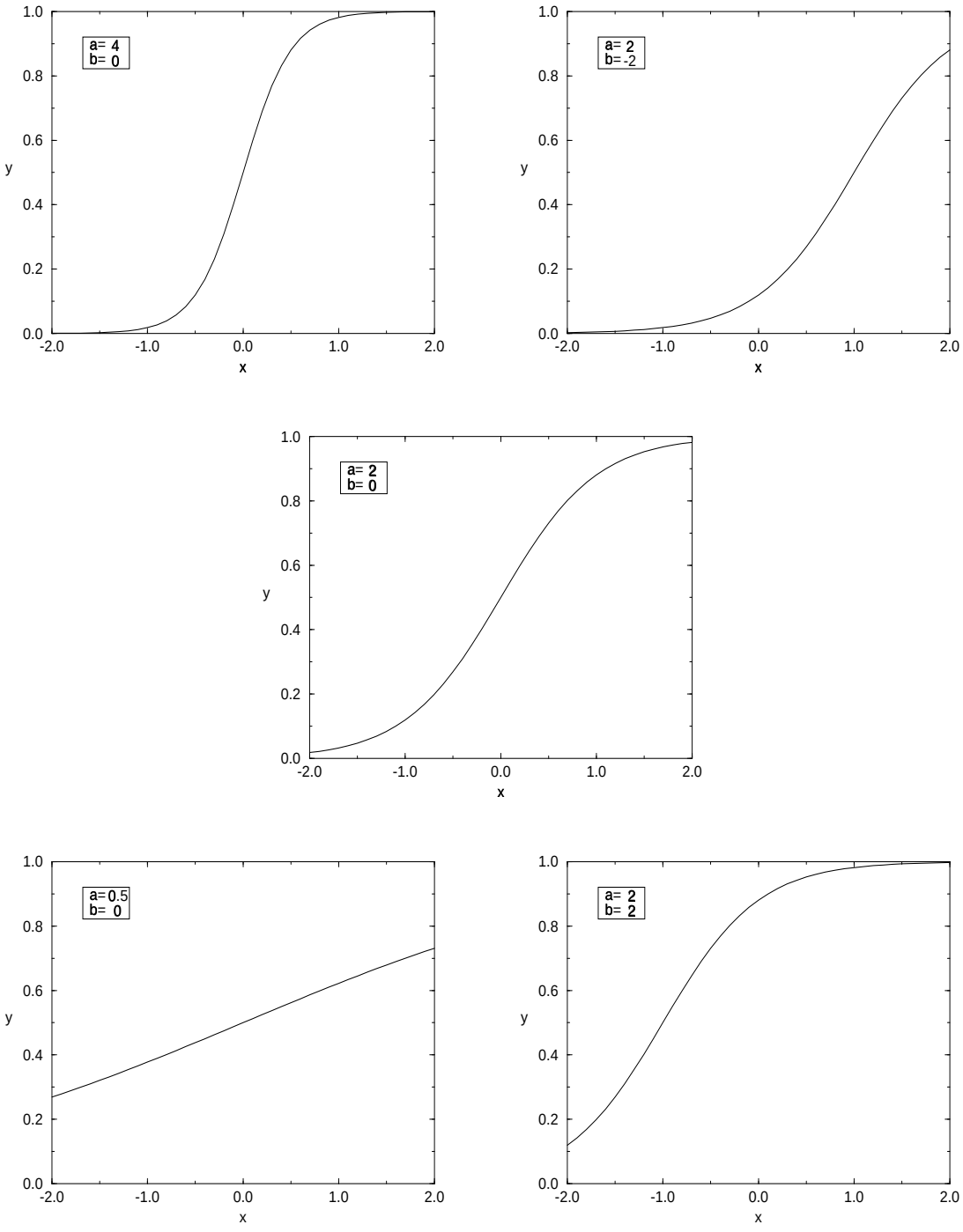


Fig. 5.3 The appearance of the function $y = \frac{1}{1+exp^{-(ax+b)}}$ for different values of a and b.

Chapter 6

The Multilayered Perceptron (MLP)

In Chapters ?? and ?? we discussed the various types of NNs and their connections to more traditional statistical methods. A specific type of NN that was discussed was the multilayered perceptron, or MLP, for short. The remainder of this book will deal with this particular type of NN.

6.1 Noise-free Data

Sometimes, but not often in the social sciences, one deals with data that have no uncertainties neither in the independent variables and nor in the dependent variables (jitter??). Such data compose nothing more than a look-up table. In other words, given some set of input variables, one may consult the data to find the corresponding unique output. Then, one question is whether one can train an MLP to encapsulate this data in such a way that the MLP will produce the correct output corresponding to some set of input variables.

Employing a MLP as a look-up table may seem far from a statistical problem; however, a great deal can be learned about the type of tables that a given kind of MLP can or cannot learn. For example, suppose we want to train a MLP to learn the data displayed in Table ??. x_1 and x_2 refer to the independent variables, and y refers to the dependent variable. In this case, they are all binary-valued.

This table is often referred to as the AND function (or rule); a little thought reveals why. Another example is the OR function, displayed in Table ??.

It is easy to show that a perceptron with two input nodes and one output node

x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	1

Table 6.1 The AND function.

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	1

Table 6.2 The OR function.

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

Table 6.3 The XOR function.

can learn these functions, if the activation function is the sign function, i.e.,

$$f(x) = sign(x) = +, if x > 0, -, if x < 0. \tag{6.1}$$

The details can be found in many texts (refs??).

The trouble begins if we attempt to teach a perceptron to learn the function known as exclusive-or (XOR), displayed in Table ??. It can be shown that no value of the weights allows for this function to be learned. Details....?

So, where are we to go at this point? The problem was solved by (Refs??). They asked what would happen if there were a "hidden layer" of nodes? Hence, the MLP.

Give more details.....

Of course, one can still do regression or classification on noise-free data.....

The question, then, arises as to what utility an MLP has in the context of noisy data, i.e., in statistics.

6.2 Noisy Data

Regression ...
Classification ...

As we mentioned previously (section??), an MLP can learn "any" function. To discuss the precise class of functions that can be learned by an MLP will take us beyond the scope of this chapter. ....

This very flexibility of MLPs makes them prone to over-fitting problems. In short, a highly nonlinear model is likely to overfit a data set, and thereby has no predictive capability at all. Although among nonlinear methods, MLPs appear to be least prone to overfitting (see the "curse of dimensionality", below), this does not imply that they cannot overfit.

An example that illustrates such overfitting is ...

Overfitting occurs because an MLP may allow for more nonlinearity than is necessary to properly fit the data. There are two quantities that determine the amount of nonlinearity in an MLP: 1) The number of hidden nodes, and 2) the magnitude (size) of the weights. These issues will be further discussed in chapter 7??.

6.3 MLPs for Regression (Supervised)

Bla bla bla....

6.4 MLPs for Classification (Supervised)

From our discussion of logistic regression, we know that ...

Just as logistic regression models class-conditional posterior probabilities, the output of an MLP can be arranged to represent class-conditional posterior probabilities as well. It has been shown that if the activation function is the logistic function, and the error function being minimized is the cross- entropy, defined as

$$S = -\sum_n (t_n \log \frac{y_n}{t_n} + (1 - t_n) \log \frac{1 - y_n}{1 - t_n}) \,, \tag{6.2}$$

then the output of the MLP is the posterior probability of class membership, given the inputs (Richard and Lippmann, 1991). In this equation, y_n is the single output of the MLP, and $t_n = 0, 1$ are the values of the dependent variable labeling the two classes. This conclusion is contingent on a training set in which the classes are represented according to their a priori probabilities; if they are not, then the outputs must be corrected for the difference (Bishop, 1996). However, many MLP applications artificially equalize the class sizes in the training set with no such corrections.

Note that an MLP with 0 hidden nodes and a logistic activation function is nothing but logistic regression, if and only if cross-entropy is minimized. This is so because logistic regression models posterior probabilities, but an MLP will model posterior probabilities only if cross-entropy is minimized. Indeed, it is the minimization of cross-entropy that yields the maximum-likelihood parameter estimates (Bishop, 1996).

6.5 Curse of dimensionality

Number of parameters in polynomial regression, grows too fast etc.

Number of parameters in LDA, i.e., the cov. matrices + means, grows too fast ....

But for MLPs, the growth is relatively slow (only linear).

6.6 NNs for Cluster Analysis (Unsupervised)

6.7 "Importance" of a Predictor

This is probably the thorniest issue that will be discussed in this book. There are many reasons for this. One reason is that often the need to know the best predictors preceds the model. For example, limited funds preclude the collection of data regarding all the variables. Of course, if one is in possession of a relatively well-proven theory that specifies the relevant variables, then ... Often, though, it is wiser to allow the data to "speak" for themselves. ....

Even if a reliable theory does exist, it is not always clear what statistical model should be tested. .... In a statistical model the importance assigned to a given variable is contingent on the model, e.g. linear regression, generalized regression, additive models, multilayered perceptron, etc.. One may even argue that there is no *unique* model that best represents the data, because it depends on what we mean by "best" and a number of other ambiguities, like when two different models are statistically equivalent, etc.. ....

Furthermore, in practice, it is very likely that some input is "important" for some domain of that input (and even of other inputs in an MLP), and not-so-important otherwise. ....

Finally, in the absence of a multiple-input model, one may invoke several 1-input, linear regression, models to provide for some ordering of the variables in terms of predictive strength. For a regression problem, i.e., with a continuous and unbounded output, this amounts to computing some linear correlation coefficient, say Pearson's ρ, between the various inputs and the output. This method has the advantage of being model-independent, but it has the disadvantage of not taking into account any nonlinearity or interactions that may be underlying the data.

6.8 Exercises

Chapter 7
Optimal Architecture

As we discussed, there are two quantities that determine the extent of nonlinearity in an MLP - the number of hidden nodes and the magnitude of the weights.

The number of hidden nodes is one quantity that gauges the complexity of the underlying function (or decision boundaries, for classification problems), i.e., the nonlinearity of the function and the complexity of the interactions between the independent variables. The magnitude of the weights is another quantity that affects the complexity of an MLP, but to a much lesser degree. By systematically varying the number of hidden nodes one effectively spans the space of "all" functions and "all" interactions (Geman, Biensenstock, and Doursat 1992; Hornik, et al. 1989). Therefore, selecting the number of hidden nodes is tantamount to specifying or selecting the underlying model in its entirety. Consequently, the "correct" number of hidden nodes is of paramount importance in any MLP development. In spite of this, many applications do not attend to this issue at all, e.g., Lenard et al. 1995, Markham and Ragsdale 1995, Warner and Misra 1996, Wilson and Hardgrave 1995.

There are a number of techniques for determining the optimal number of hidden nodes. These will be discussed next. Additionally, there are methods that are designed to (partially) avoid the problem of what the optimal number of hidden nodes should be. An example is the weight-decay method, also described below.

7.1 Bootstrapping

The standard and well-known technique of boostrapping predates neural networks (Refs??). It is a method for estimating the generalization error. It is, however, one of the most popular ways by which the optimal number of hidden nodes can be determined.

The data is randomly divided into two sets - a training and a validation set. The training set is employed to estimate the parameters of the MLP, and the validation set is utilized to gauge the predictive performance of the MLP. A common error occurs when the the number of hidden nodes is selected to be that which minimizes

the validation error. This is a mistake, because although the network has not overfit the training set, it has overfit the validation set!

To avoid this type of overfitting, the data is randomly divided again, and again, some number of times, and the training and validation results are averaged over the different sets. Of course, when one can average some number of entities, one can compute the standard error for the mean as well, and this can serve as a measure of confidence in the means.

Give some simple example of how this works.

7.2 Cross-Validation

This is another method for determining the optimal number of hidden nodes based on re-sampling. It is similar to Bootstrapping in that the data is resampled many times. However, instead of partitioning the data (into training and validation) say, k times, one divides the data in k subsets, using k-1 of the subsets for training and the remaining subset for validation.

Give a simple example of this, too.

7.3 Weight-decay

In the method of weight decay, the question of the optimal number of hidden nodes goes away! How can this be? The answer becomes evident when one notes that the reason overfitting can occur is that the magnitude of the parameters (weights) can become exceedingly large. With large weights any nonlinear activation function becomes highly nonlinear. For example, for large weights the logistic function (Figure ??) can approach a step-function, which is an example of a highly nonlinear function. Because the network is a (nonlinear) combination of a whole bunch of these activation functions, the MLP can becomes an extremely nonlinear function thereby overfitting data.

Therefore, one way to avoid overfitting is to avoid the weights from becoming too large. Exactly how large is too large is a thorny issue that will addressed below. To keep the weights from growing too large during training, one adds a weight-decay term to the standard error function (Eq. ??):

$$E = \frac{1}{N?}\sum_{i=1}^{N}[t_i - y_i(\omega)]^2 + \frac{\alpha}{N}\sum_{j=1}^{No.of parameters}\omega_j^2. \qquad (7.1)$$

The parameter α determines how large the weights can get; a large value of α can keep the weights small, and vice versa.

More on this.....

7.4 Bias vs. variance

The performance of the MLP on the training set is positively biased, yet many studies compute only this measure. For a less- biased estimate, performance must be gauged on the validation set. For a completely unbiased measure of performance, a "third" data set - often called the test set - is required. However, the division of the data set into three sets reduces the size of each of the training, validation, and test sets. This, in turn, increases the variance of the estimates. It has been shown that a compromise to this "bias vs. variance dilemma" (Geman, et al. 1992) is offered by cross-validation.

7.5 Local Minima

An important issue in the training of MLPs is that of the local minima of the error function. Most learning algorithms (i.e. parameter estimation techniques) are iterative procedures where a set of randomly selected weights are slowly varied in an attempt to minimize the error function. Given that equation 1 is nonlinear in the weights, frequently the learning algorithm gets trapped in a local minimum of the error function. Such an MLP does not correctly represent the underlying structure of the data. The simplest way of dealing with this problem is to repeat the entire learning phase from a different random set of initial weights.

7.6 Autocorrelation?

7.7 Exercises

Chapter 8
MLP for Regression

8.1 Generalities

Bla bla bla...

8.2 Bootstrapping

Bla bla bla...

8.3 Application: Simulated Data

Same as the classification example ....
Do example with many hidden nodes to show overfitting.....
Plot the distribution of the residues....
Illustrate Bootstrapping.....

8.4 Application: Television Viewing and High School Mathematics Achievement, revisited

This section re-examines the relationship between mathematics achievement and television viewing (section ??), but this time a feed-forward NN is employed as a nonlinear model. We shall then compare the findings with those of section ??. A similar curvilinear relationship is found, independent of viewer characteristics, parental background, parental involvement, and leisure activities, with a peak at about one hour of viewing, and persistent upon the inclusion of statistical errors. It is further shown that for low-ability students the curvilinearity is replaced with an entirely positive correlation across all hours of television viewing (Paik, 1998).

Bootstrapping can be employed to arrive at the optimal number of hidden nodes. First, the data are randomly divided into two sets - a training set and a validation

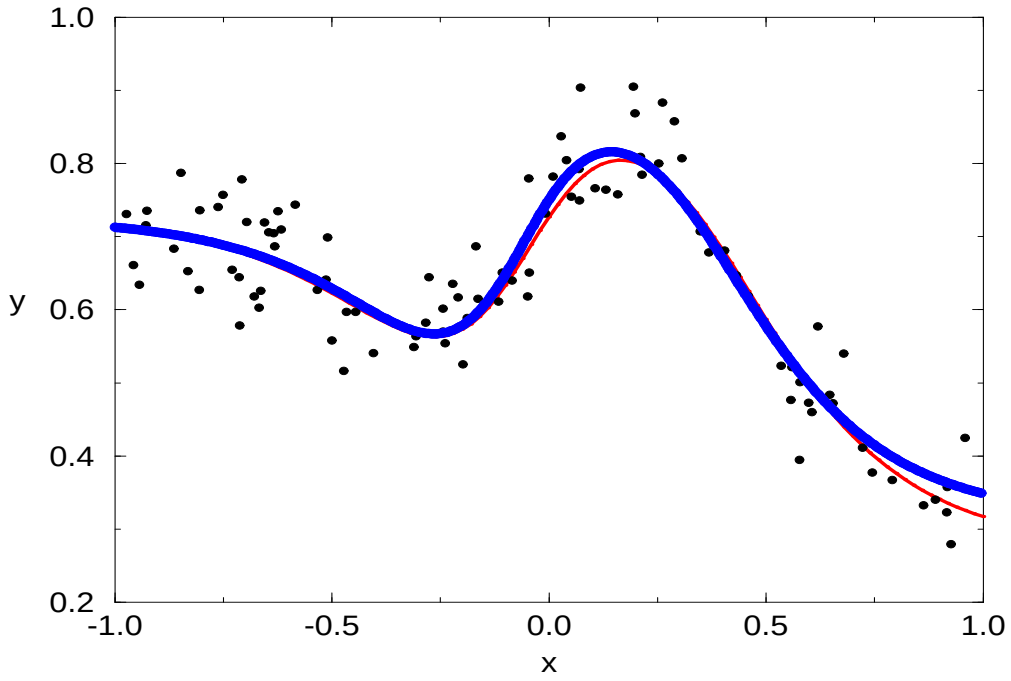


Fig. 8.1 The data (black), the underlying function (red), and the fit according to an MLP with 2 hidden nodes (blue).

set. An NN with 0 hidden nodes is trained with the training set and its performance is gauged on the validation set. The number of hidden nodes is then incremented and the procedure repeated until the validation error begins to rise. In this way one arrives at the number of hidden nodes that precludes overfitting the training set. As mentioned previously, this procedure leads to an NN that overfits the validation set. To preclude overfitting the validation set, the original data set is divided again but with a different random partitioning into a training and a validation set, and the entire procedure is repeated again. The validation errors over the different random sets can then be employed to compute an average and a confidence interval for the validation performance measures. The optimal number of hidden nodes is the value beyond which the average validation error begins to rise.

The output node of the network is taken to be the mathematics scores. First, a network with only one input node - the amount of television viewing - is trained and tested. Then the number of input nodes is increased to two, three, etc., with the additional nodes representing gender, SES, ability, etc.. Finally, for each trained network the input node corresponding to television viewing is varied from 0 (0 hours

of viewing per day) to 6 (5 or more hours of viewing per day), in 0.1 increments, while controlling the other input variables, and recording the output values (math. scores). The standard errors are also computed at each viewing hour.

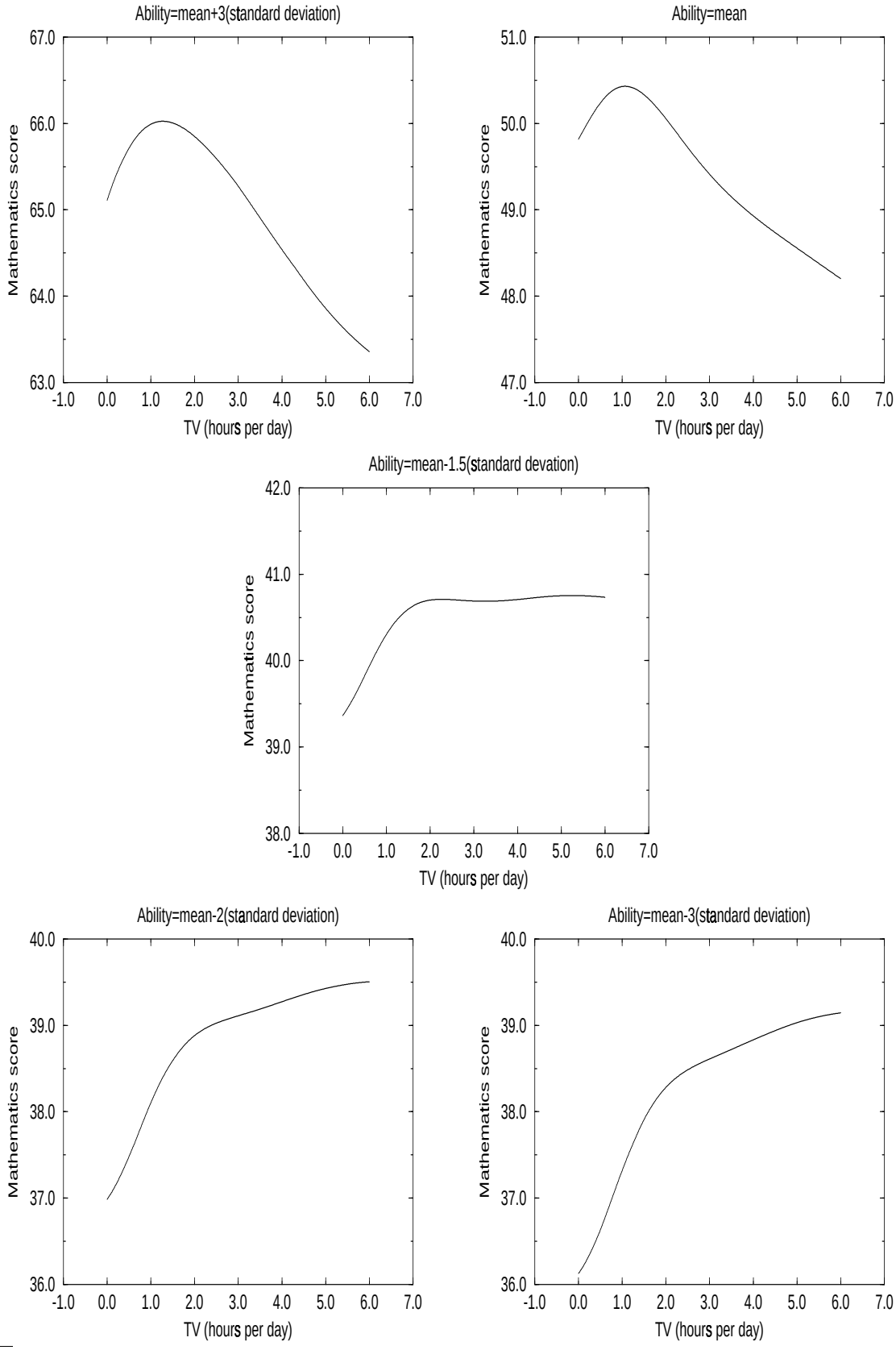
The analysis indicates that there is a statistically significant curvilinear relationship between high school mathematics performance and the amount of television viewing, even when a host of viewer characteristics, parent-related, and leisure-related variables are controlled.

By fixing the values of the input nodes corresponding to gender and SES, one can obtain similar plots but for these variables controlled:

Gender and SES curves.

Additionally, it is shown that for low ability students, the correlation between mathematics performance and television viewing is positive across all viewing hours.

8.5 Exercises


Ability=mean+3(standard deviation)
Mathematics score
TV (hours per day)
Ability=mean
Mathematics score
TV (hours per day)
Ability=mean-1.5(standard devation)
Mathematics score
TV (hours per day)
Ability=mean-2(standard deviation)
Mathematics score
TV (hours per day)
Ability=mean-3(standard deviation)
Mathematics score
TV (hours per day)

Chapter 9
MLP for classification

9.1 Binary

Illustrate with examples, probability density, Bayes, Gaussian classifier...

9.2 C Classes

Correlation matrix becomes important. Illustrate with Gaussian.

9.3 Application: Simulated Data

Before we place too much faith in the capabilities of an MLP, why do we not test it with a problem whose solution we know with certainty? In other words, let us create data from a known distribution with a known boundary demarking the, say 2, classes.

A data set was generated with two independent variables, x_1, and x_2, ranging from -1 to +1, and one dependent variable whose 0/1 values label two groups. This was done such that the decision boundary between the two groups, i.e., the inverted "Mexican hat" with the filled circles in Figure ??, corresponds to that of a network with 4 hidden nodes. Indeed, prior to the addition of noise to the data, a 4-hidden-node network can learn this boundary with zero error. The addition of some gaussian noise with a standard deviation of 0.2 produces the data appearing in Figure ?? with the lower-pointing filled triangles representing one group and the upper-pointing triangles representing the second group. The training set contained 300 cases and the validation set 200. The empty circles in Figures 1a-1c outline the decision boundaries produced by several networks with different number of hidden nodes.

We can see that the MLP with 4 hidden nodes correctly learns the boundary, while the one with 15 hidden nodes overfits it; in addition to the excessive meandering of the MLP's fit about the underlying boundary, note the false boundaries

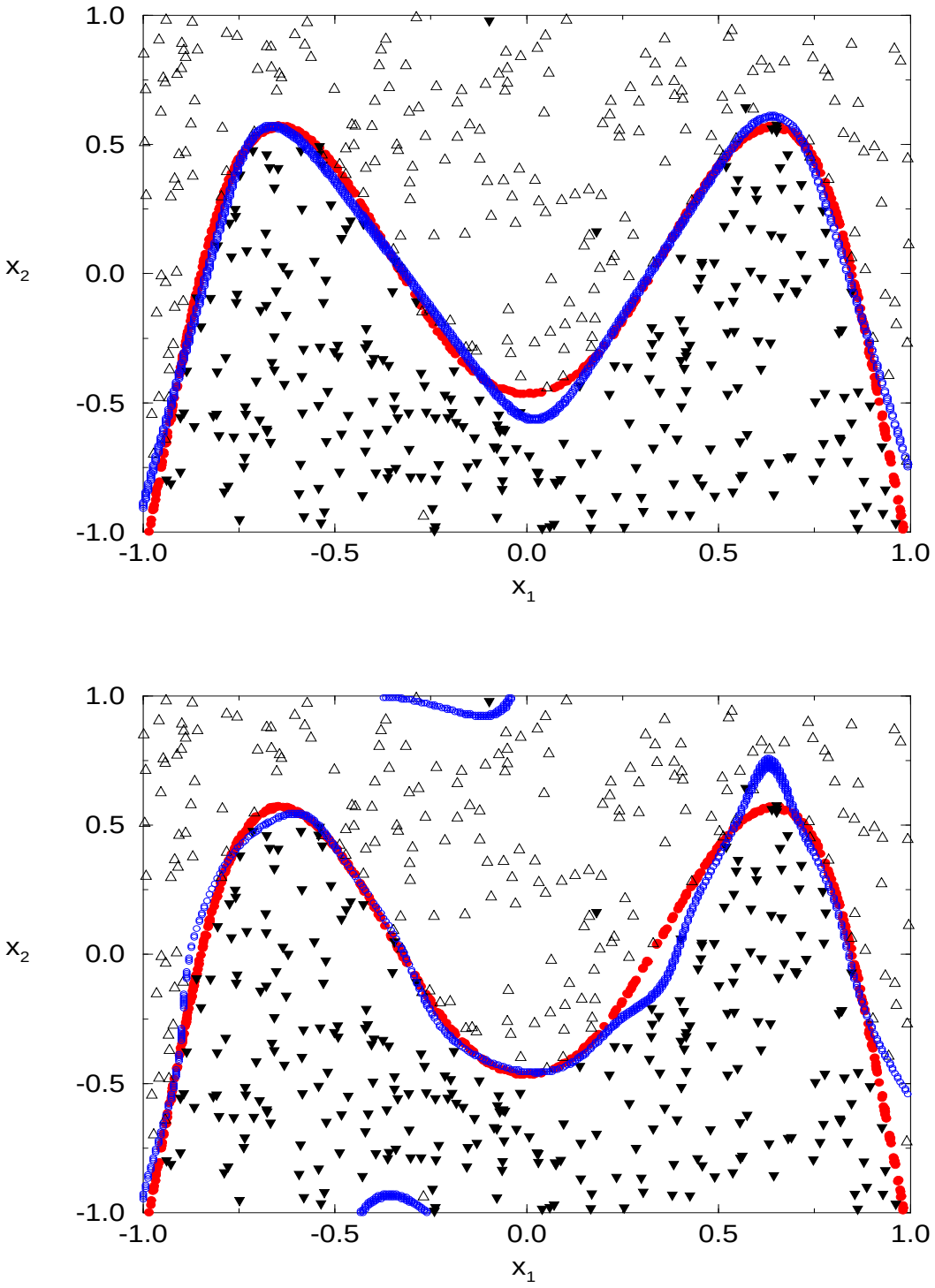


Fig. 9.1 The data (black and white triangles) , the underlying boundary (red), and the estimated boundary (blue) according to an MLP with 4 (top) and 15 (bottom) hidden nodes.

at the top and the bottom of Figure ?? as the network attempts to create separate decision regions for two individual data points.

9.4 Application: Television Extreme Viewers and Nonviewers, Revisited

Figure ?? shows the neural network's percent efficiency in the prediction of nonviewers and extreme viewers, separately. LDA results are also included for comparison, although they will not be discussed since we have shown that it does not represent reality as faithfully as the NN does. We conclude that the strongest attributes of extreme viewers are family and activity- related variables combined (case (23) at 88%), whereas nonviewers are best characterized by their demographics (case (1) at 78%). It is interesting, and surprising, that although demographic variables also moderately characterize extreme viewers (case (1) at 71%), family-related and activity-related variables, separately or combined, do not characterize nonviewers at all (case (2), (3), or (23) at 52% to 67%).

Update/revise these results.

PLDA performs comparably to the linear NN (i.e., with 0 hidden nodes, or logistic regression). This is somewhat surprising given the data's violation of the explicit assumptions of normality and homoelasticity invoked in LDA. However, the robustness of LDA under violations of its assumptions is well- known (Lachenbruch 1975).

9.5 Application: Predicting High School Delinquency, Revisited

9.6 Exercises

Bibliography

A. C. Nielsen and Company (1998) *1997-1998 Nielsen Report on television.* New York: Author

Amit, D. J. (1992) *Modeling brain function: The world of attractor neural networks.* Cambridge, England: Cambridge University Press.

Anderer, P., Saletu, B., Kloppel, B., Semlitsch, H. V., and Werner, H. (1994) "Discrimination between demented patients and normals based on topographic EEG slow wave activity: Comparison between z statistics, discriminant analysis and artificial neural network classifiers", *Electroencephalography and Clinical Neurophysiology* **91**,108.

Anderson, D. R., and Field, D. E. (1991) "Online and offline assessment of the television audience." In J. Bryant, and D. Zillmann (Eds.), *Responding to the screen: Reception and reaction processes (pp. 199-216).* Hillsdale, NJ: Lawrence Erlbaum.

Artysushkin, V. F., Belyayev, A. V., Sandler, Y. M., and Serveyev, V. M. (1990) "Neural network ensembles as models of interdependence in collective behavior", *Mathematical Social Sciences* **19**, 167.

Azari, N. P., Pietrini, P., Horwitz, B., and Pettigrew, K. D. (1993) "Individual differences in cerebral metabolic patterns during pharmacotherapy in obsessive-compulsive disorder: A multiple regression/discriminant analysis of positron emission tomographic data", *Biological Psychiatry* **34**, 798.

Bates, D. M., and Watts, D. G. (1988) *Nonlinear regression analysis and its applications.* New York, NY: John Wiley and Sons, Inc.

Becktel, R. P., Achelpohl, C., and Akers, R. (1972) "Correlates between observed behavior and questionnaire responses on television viewing", In E. A. Rubinstein, G. A. Comstock, and J. P. Murray (Eds.), *Television and social behavior: Vol. 4. Television in day-to-day life: Patterns of use (pp. 274-344).* Washington, DC: Government Printing Office.

Beentjes, J. W. M., and van der Voort, T. J .A. (1988). "Television's impact on children's reading skills: A review of research", *Reading Research Quarterly* **23**, 389.

Bertalanffy, L. (1968) *General system theory: Foundations development applications.* New York: Braziller.

Bishop, C. M. (1996) *Neural networks for pattern recognition.* Oxford: Clarendon Press.

Boone, S. L. (1991) "Aggression in African-American boys: A discriminant analysis", *Genetic, Social, and General Psychology Monographs* **117**, 203.

Buckley, W. (1968) "Society as a complex adaptive system", In Walter Buckley (ed.), *Modern Systems Research for the Behavioral Scientist (pp. 490-513).* Chicago: Aldine.

Camilli, G. (1990) "The test of homogeneity for 2 x 2 contingency tables: A review of and some personal opinions on the controversy", *Psychological Bulletin* **108**, 135.

Carnevalli, P., and Patarnello, S.(1987) "Exhaustive Thermodynamical Analysis of Boolean Learning Networks", *Europhysics Letters* **4**, 1199.

Cheng, B., and Titterington, D. M. (1994), "Neural Networks: A Review from a Statistical Perspective", *Statistical Science* **9**, 2.

Cherkassky, V., and Mulier, F. (1994) "Statistical and neural network techniques for non-parametric regression", *Lecture notes in statistics* **89**, 383.

Cherkassky, V., Friedman, J. H., and Wechsler, H., eds. (1994) *From Statistics to Neural Networks: Theory and Pattern Recognition Applications.* Berlin: Springer-Verlag.

Christensen, L., and Duncan, K. (1995) "Distinguishing depressed from nondepressed individuals using energy and psychosocial variables", *Journal of Consulting and Clinical Psychology* **63**, 495.

Churchland, P. S., and Sejnowski, T. J. (1993) *The computational brain.* Cambridge, MA: The MIT Press.

Collins, E., Ghosh, S., and Scofield, S. (1988) *Risk Analysis: DARPA Neural Network Study.* AFCEA International Press.

Collins, W. A. (1982) "Cognitive processing in television viewing", pp.9-23 in D. Pearl, L. Bouthilet, and J. Lazar (eds.), *Television and behavior: Ten years of scientific progress and implications for the eighties: Vol. 2. Technical reviews.* Rockville, MD: National Institute of Mental Health.

Comstock, G., and Paik, H. (1991) *Television and the American child.* San Diego, CA: Academic Press, Inc. Dammers, E. (1993) "Measurement in the ex post evaluation of forecasts", *Quality and Quantity* **27**, 31.

Davis, J. A., and Smith, T. W. (1990) *General Social Surveys, 1972-1990.* [Electronic data tape]. NORC ed. Chicago: National Opinion Research Center [Producer]; Storrs, CT: The Roper Center for Public Opinion Research, University of Connecticut [Distributor].

Doolittle, M. H. (1888) "Association ratios", *Bulletin of the Philosophical Society of Washington* **10**, 83.

Draper, N. R., and Smith, H. (1981) *Applied regression analysis.* New York: John Wiley and Sons.

Dunteman, G. H. (1984) *Introduction to multi variate analysis.* Beverly Hills, CA: Sage Publications.

Edgar, P. (1977) "Families without television", *Journal of Communication* **27**, 73.

Evans, T. David., Francis T. Cullen, Burton, V. S. Jr., Dunaway, R. G., Payne, G. L., and Kethineni, S. R. (1996) "Religion, social bonds, and delinquency", *Deviant Behavior: An Interdisciplinary Journal* **17**, 43.

Famularo, R., Fenton, T., Kinscherff, R., Barnum, R., Bolduc, S., and Bunschaft, D. (1992) "Differences in neuropsychological and academic achievement between adolescent delinquents and status offenders", *American Journal of Psychiatry* **149**, 1252.

Fetler, M. (1984) "Television viewing and school achievement", *Journal of Communication* **34**, 104.

Fienberg, S. E. (1980) *The analysis of cross- classified categorical data (2nd ed.)* Cambridge, MA: MIT Press.

Fosarelli, P. (1986) "In my opinion... Advocacy for children's appropriate viewing of television: What can we do?", *Children's Health Care* **15**, 79.

Gallant, A. R. (1987) *Nonlinear Statistical Models.* NY: Wiley.

Garson, D. G. (1991) "A comparison of neural network and expert systems algorithms with common multi variate procedures for analysis of social science data", *Social*

Science Computer Review **9**, 399.

Gaddy, G. D. (1986) "Television's impact on high school achievement", *Public Opinion Quarterly* **50**, 340.

Geman, S., Biensenstock, E., and Doursat, R. (1992) "Neural networks and the bias/variance dilemma", *Neural Computation* **4**, 1.

Gerbner, G., Gross, L., Jackson-Beeck, J., Jeffries- Fox, S., and Signorielli, N. (1978) "Cultural indicators: Violence profile No. 9", *Journal of Communication* **28**, 176.

Giddens, A. (1982) *Profiles and Critiques in social theory.* Berkeley, CA: University of California Press.

Glick, N. (1978) "Additive estimators for probabilities of correct classification", *Pattern Recognition* **10**, 211.

Goodman, L. A., and Kruskal, W. H. (1954) "Measures of association for cross classifications", *American Statistical Association Journal* **49**, 723.

Goodman, L. A., and Kruskal, W. (1959) "Measures of association for cross classifications", *Journal of the American Statistical Association* **54**, 123.

Greenstein, J. (1954) "Effects of television upon elementary school grades", *Journal of Educational Research* **48**, 161.

Gutfreund, H., and Toulouse, G. (E's) (1994) *Biology and computation. Advanced series in neuroscience: Vol. 3.* Singapore: World Scientific Publishing Company.

Haertel, E. H., and Wiley, D. E. (1978) *Social and economic differences in elementary school achievement.* Chicago: ML-Group for Policy Studies in Education, CEMREL, Inc.

Haber, M. (1990) "Comments on "the test of homogeneity for 2 x 2 contingency tables: A review of and some personal opinions on the controversy" by G. Camilli", *Psychological Bulletin* **108**, 146.

Hammond, S. M., and Lienert, G. A.(1995) "Modified Phi Correlation for the multi variate analysis of ordinally scaled variables", *Educational and Psychological Measurement* **55**, 225.

Hand, D. J. (1981) *Discrimination and Classification.* NY: Wiley.

Hardgrave, B. C., Wilson, R. L., and Walstrom, K. A. (1994) "Predicting graduate student success: a comparison of neural networks and traditional techniques", *Computers and Operation Research* **21**, 249.

Hays, W. L. (1973) *Statistics for the social sciences (2nd ed.)* New York, NY: Holt, Rinehart and Winston.

Hill, T., Marquez, L., O'Connor, M., and Remus, W. (1994) "Artificial neural network models for forecasting and decision making", *International Journal of Forecasting* **10**, 5.

Hirsch, P. M. (1980) "The "Scary world" of the nonviewer and other anomalies: A reanalysis of Gerbner et al.,'s findings on cultivation analysis, Part I", *Communication Research* **7**, 403.

Hornik, K., Stinchcombe, M., and White, H. (1989) "Multilayer feedforward networks are universal approximators", *Neural Networks* **4**, 251.

Huberty, C. A. (1994) *Applied discriminant analysis.* New York, NY: John Wiley and Sons, Inc.

Huntley, D. G. (1991) "Neural nets: An approach to the forecasting of time series", *Social Science Computer Review* **9**, 27.

Huston, A., and Wright, J. C. (1989) "The forms of television and the child viewer", In Comstock, G. (Ed.), *Public communication and behavior (Vol. 2, pp. 103-159).* New York: Academic Press.

Jackson-Beeck, M. (1977) "The nonviewers: who are they?", *Journal of Communication*

27, 65.

Keith, T. Z., Reimers, T. M., Fehrmann, P. G., Pottebaum, S. M., and Aubey, L. W. (1986) "Parental involvement, homework, and TV time: Direct and indirect effects on high school achievement", *Journal of Educational Psychology* **78**, 373.

Kendall-Tackett, K. A. (1996) "The effects of neglect on academic achievement and disciplinary problems: A developmental perspective", *Child Abuse and Neglect* **20**, 161.

Klecka, W. R. (1980) *Discriminant analysis.* Newbury Park, CA: Sage Publications.

Kubey, R. W., and Czikszentmihalyi, M. (1990) *Television and the quality of life. How viewing shapes everyday experience.* Hillsdale, NJ: Erlbaum.

Lachenbruch, P. A. (1975) *Discriminant analysis.* New York, NY: Hafner Press.

Lachenbruch, P. A., and Mickey, M. A. (1968) "Estimation of error rates in discriminant analysis", *Technometrics* **10**, 1.

Lapointe, A. E., Mead, N. A., and Askew, J. M. (1992) *Learning mathematics.* Princeton, N.J.: Educational Testing Service.

Lee, V. E., and Smith, J. B. (1994) *High school restructuring and student achievement. A new study finds strong links. Issue report no. 7.* (ERIC Document Reproduction Service No. ED 326-565)

Lenard, M. J., Alam, P., and Madey, G. R. (1995) "The application of neural networks and a qualitative response model to the auditor's going concern uncertainty decision", *Decision Sciences* **26**, 209.

Leven, S. J., and Levine, D. S. (1996) "Multiattribute decision making in context: A dynamic neural network methodology", *Cognitive Science* **20**, 271.

Lull, J. (1990) "Families' social uses of television as extensions of the household", In J. Bryant (Ed.), *Television and the American Family (pp. 59-72)* Hillsdale, NJ: Erlbaum.

Macy, M. (1996) "Natural selection and social learning in prisoner's dilemma: Coadaptation with genetic algorithms and artificial neural networks", *Sociological Methods & Research* **25**, 103.

Markham, I. S., and Ragsdale, C. T. (1995) "Combining neural networks and statistical predictions to solve the classification problem in discriminant analysis", *Decision Sciences* **26**, 229.

Marzban, C. (1998). "Scalar measures of performance in rare-event situations", *Weather and Forecasting* **11**, 13.

Marzban, C., Paik, H., and Stumpf, G. (1997) "Neural networks vs. gaussian discriminant analysis", *AI Applications* **11**, 1.

Marzban, C., and Stumpf, G. (1995) "A neural network for tornado prediction based on Doppler radar-derived attributes", it Journal of Applied Meteorology **35**, 617.

Masters, T. (1993) *Practical neural network recipes in C++.* Academic Press, Inc.

McCullagh, P., and Nelder, J. A. (1989) *Generalized Linear Models, 2nd ed.* London: Chapman & Hall.

McLachlan, G. J. (1992) *Discriminant analysis and statistical pattern recognition.* New York, NY: A Wiley- Interscience Publication.

Meraviglia, C. (1996) "Models of representation of social mobility and inequality systems. A neural network approach", *Quality and Quantity* **30**, 231.

Michie, D., Spiegelhalter, D. J. and Taylor, C. C. (1994) *Machine Learning, Neural and Statistical Classification.* England: Ellis Horwood.

Moore, D. S., and McCabe, G. P. (1989) *Introduction to the Practice of Statistics.* NY: W.H. Freeman.

Mller, B., and Reinhardt, J. (1990) *Neural networks; an introduction. Physics of Neural Networks Series.* Berlin, New York: Springer-Verlag.

Murphy, A.H., and Winkler, R.L. (1992) "Diagnostic verification of probability forecasts", *International Journal of Forecasting* **7**, 435.

Myers, R. H. (1986) *Classical and Modern Regression with Applications.* Boston: Duxbury Press.

National Institute of Mental Health. (1982) *Television and behavior: Ten years of scientific progress and implications for the eighties* (DHHS Publication No. ADM 82- 1195). Washington, DC: U.S. Government Printing Office.

National Opinion Research Center. (1980) *High School and Beyond information for users: Base year (1980) data.* Chicago: Author.

Nordbotten, S. (1997) "Models of complex human screening and correcting of social data", *Computers in Human Behavior* **13**, 487.

Ott, L., Larson, R. F., and Mendenhall, W. (1983) *Statistics: A tool for the social sciences (3rd ed.)* Boston, MA: Duxbury Press.

Paik, H. (1998) "The Effect of Prior Probability On Skill in Two-Group Discriminant Analysis", *Quality and Quantity* **32**, 1.

Paik, H. (1998) "Television Viewing and High School Mathematics Achievement: A Neural Network Analysis", *Quality and Quantity*, in press.

Paik, H., and Marzban, C. (1995) "Predicting television extreme viewers and nonviewers: A neural network analysis", *Human Communication Research* **22**, 284.

Paik, H., and Comstock, G. (1994) "The effects of television violence on antisocial behavior: A meta-analysis", *Communication Research* **21**, 516.

Parshall, C. G., and Kromrey, J. D. (1996) "Tests of independence in contingency tables with small samples: A comparison of statistical power", *Educational and Psychological Measurement* **56**, 26.

Peirce, C. S. (1884) "The numerical measure of the success of predictions [Letter to the editor]", *Science* **4**, 453.

Potter, W. J. (1987) "Does television viewing hinder academic achievement among adolescents?", *Human Communication Research* **14**, 27.

Reby, D., Lek, S., Dimpoulos, I., Joachim, J. *et al.* "Artificial neural networks as a classification method in the behavioural sciences", *Behavioural Processes* **40**, 35.

Richard, M. D., and Lippmann, R. P. (1991) "Neural network classifiers estimate Bayesian a-posteriori probabilities", *Neural Computation* **3**, 461.

Ridley-Johnson, R., Cooper, H., and Chance, J. (1983) "The relation of children's television viewing to school achievement and I.Q",*Journal of Educational Research* **76**, 294.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks.* Cambridge: Cambridge University Press.

Salomon, G. (1994) *Interaction of media, cognition and learning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Sarle, W. S. (1994) "Neural Networks and Statistical Models in SAS Institute Inc." *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc., 1538-1550 ftp://ftp.sas.com/pub/neural/neural1.ps.)

Sarle, W. S. (1994) "Neural Network Implementation in SAS Software in SAS Institute Inc." *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc., 1551-1573.

Schrodt, P. A. (1991) "Prediction of interstate conflict outcomes using a neural network", *Social Science Computer Review* **9**, 359.

Smith, R. (1986) "Television addiction", In J. Bryant and D. Zillmann (Eds.), *Perspectives on Media Effects (pp. 109-128)* Hillsdale, NJ: Lawrence Erlbaum Associates.

Stevens, J. P. (1996) *Applied multivariate statistics for the social sciences.* Mahwah, NJ:

Lawrence Erlbaum.

Stone, M. (1974) "Cross-validatory choice and assessment of statistical predictions", *Journal of the Royal Statistical Society B* **36**, 111.

Stone, M. (1974) "Cross-validatory choice and assessment of statistical predictions", *Journal of the Royal Statistical Society B* **36**, 111.

Tankard, J. W., Jr., and Harris, M. C. (1980) "A discriminant analysis of television viewers and nonviewers", *Journal of Broadcasting* **24**, 399.

U.S. Department of Education, National Center for Education Statistics (1992) "National Education Longitudinal Study, 1988: First Follow-up (1990)." [Student Data] [Computer file]. U.S. Department of Education, Office of Educational Research and Improvement [producer], 1992. Ann Arbor, MI: Inter- university Consortium Political and Social Research [distributor], 1992.

Van Nelson, C., and Neff, K. J. (1990) "Comparing and contrasting neural net solutions to classical statistical solutions", *Paper presented at the Annual Meeting of the Midwestern Educational Research Association (Chicago, IL)* (ERIC Document Reproduction Service No. ED 326-577)

Wahba, G., and Wold, S. (1975) "A completely automatic French curve: fitting spline functions by cross-validation", *Communications in Statistics, Series A* **4**, 1.

Warner, B., and Misra, M. (1996) "Understanding neural networks as statistical tools", *The American Statistician* **50**, 284.

Weisberg, S. (1985) *Applied Linear Regression.* NY: Wiley

Westley, B. H., and Mobius, J. B. (1960) "A closer look at the non-television household", *Journal of Broadcasting* **4**, 164.

Wilcox, R. R. (1996) *Statistics for the social sciences.* San Diego, CA: Academic Press.

Wilson, R. L., and Hardgrave, B. C. (1995) "Predicting graduate student success in an MBA program regression versus classification", *Educational Psychological Measurement* **55**, 186.

White, H., and Gallant, A. R. (1992) "On Learning the Derivatives of an Unknown Mapping with Multilayer Feedforward Networks", *Neural Networks* **5**, 129.

White, H. (1990) "Connectionist Nonparametric Regression: Multilayer Feedforward Networks Can Learn Arbitrary Mappings", *Neural Networks* **3**, 535.

Woelfel, J. (1993) "Artificial neural networks in policy research: A current assessment", *Journal of Communication* **43**, 63.

Woelfel, J. (1993) "Cognitive processes and communication networks: A general theory", In W. D., Richards and G. A., Barnett (Eds.), *Progress in communication sciences: Volume XII (pp.21-42)* New Jersey: Ablex Publishing Co.

Woelfel, J., Richards, W. D., and Stoyanoff, N. J. (1993) "Conversational networks", In W. D., Richards and G. A., Barnett (Eds.), *Progress in communication sciences: Volume XII (pp.223-246)* New Jersey: Ablex Publishing Co.

Zillmann, D. (1982) "Television viewing and arousal", pp. 53-67 in D. Pearl, L. Bouthilet, and J. Lazar (eds.), *Television and behavior: Ten years of scientific inquiry and implications for the eighties: Vol. 2. Technical reviews.* Washington, DC: U.S. Government Printing Office.